Contents lists available at ScienceDirect

Solar Energy

journal homepage: www.elsevier.com/locate/solener

Verification of deterministic solar forecasts[☆]

Dazhi Yang[a,*], Stefano Alessandrini[b], Javier Antonanzas[c], Fernando Antonanzas-Torres[c], Viorel Badescu[d], Hans Georg Beyer[e], Robert Blaga[f], John Boland[g], Jamie M. Bright[h], Carlos F.M. Coimbra[i], Mathieu David[j], Âzeddine Frimane[k], Christian A. Gueymard[l], Tao Hong[m], Merlinde J. Kay[n], Sven Killinger[o], Jan Kleissl[i], Philippe Lauret[j], Elke Lorenz[o], Dennis van der Meer[p], Marius Paulescu[f], Richard Perez[q], Oscar Perpiñán-Lamigueiro[r], Ian Marius Peters[s], Gordon Reikard[t], David Renné[u], Yves-Marie Saint-Drenan[v], Yong Shuai[w], Ruben Urraca[c], Hadrien Verbois[h], Frank Vignola[x], Cyril Voyant[j], Jie Zhang[y]

[a] Singapore Institute of Manufacturing Technology, Agency for Science, Technology and Research, Singapore
[b] Research Application Laboratory, National Center for Atmospheric Research, Boulder, CO, USA
[c] Deparment of Mechanical Engineering, University of La Rioja, Logrono, Spain
[d] Candida Oancea Institute, Polytechnic University of Bucharest, Bucharest, Romania
[e] Faculty of Science and Technology, University of the Faroe Islands, Tórshavn, Faroe Islands
[f] Faculty of Physics, West University of Timisoara, Timisoara, Romania
[g] Centre for Industrial and Applied Mathematics, University of South Australia, Mawson Lakes, SA, Australia
[h] Solar Energy Research Institute of Singapore, National University of Singapore, Singapore
[i] Department of Mechanical and Aerospace Engineering, University of California, San Diego, CA, USA
[j] PIMENT Laboratory, University of La Reunion, Reunion, France
[k] Faculty of Science, Ibn Tofail University, Kenitra, Morocco
[l] Solar Consulting Services, Colebrook, NH, USA
[m] Department of Systems Engineering and Engineering Management, University of North Carolina at Charlotte, Charlotte, NC, USA
[n] School of Photovoltaic and Renewable Energy Engineering, University of New South Wales, Sydney, NSW, Australia
[o] Fraunhofer Institute for Solar Energy Systems ISE, Freiburg, Germany
[p] Department of Civil and Industrial Engineering, Uppsala University, Uppsala, Sweden
[q] Atmospheric Sciences Research Center, University at Albany, SUNY, Albany, NY, USA
[r] School of Engineering and Industrial Design, Polytechnic University of Madrid, Madrid, Spain
[s] Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[t] Statistics Department, U.S. Cellular, Chicago, IL, USA
[u] Dave Renné Renewables, Boulder, CO, USA
[v] MINES ParisTech, PSL Research University, Sophia Antipolis, France
[w] School of Energy Science and Engineering, Harbin Institute of Technology, Harbin, Heilongjiang, China
[x] Material Science Institute, University of Oregon, Eugene, OR, USA
[y] Department of Mechanical Engineering, University of Texas at Dallas, Richardson, TX, USA

ARTICLE INFO

Keywords:
Solar forecasting
Measure-oriented forecast verification
Distribution-oriented forecast verification
Skill score
Combination of climatology and persistence

ABSTRACT

The field of energy forecasting has attracted many researchers from different fields (e.g., meteorology, data sciences, mechanical or electrical engineering) over the last decade. Solar forecasting is a fast-growing subdomain of energy forecasting. Despite several previous attempts, the methods and measures used for verification of deterministic (also known as single-valued or point) solar forecasts are still far from being standardized, making forecast analysis and comparison difficult.

To analyze and compare solar forecasts, the well-established Murphy–Winkler framework for distribution-oriented forecast verification is recommended as a standard practice. This framework examines aspects of forecast quality, such as reliability, resolution, association, or discrimination, and analyzes the joint distribution of forecasts and observations, which contains all time-independent information relevant to verification. To verify forecasts, one can use any graphical display or mathematical/statistical measure to provide insights and

[☆] All co-authors are listed in alphabetical order.

[*] Corresponding author.

E-mail address: yangdazhi.nus@gmail.com (D. Yang).

https://doi.org/10.1016/j.solener.2020.04.019
Received 14 November 2019; Received in revised form 11 February 2020; Accepted 7 April 2020
Available online 18 May 2020

summarize the aspects of forecast quality. The majority of graphical methods and accuracy measures known to solar forecasters are specific methods under this general framework.

Additionally, measuring the overall skillfulness of forecasters is also of general interest. The use of the root mean square error (RMSE) skill score based on the optimal convex combination of climatology and persistence methods is highly recommended. By standardizing the accuracy measure and reference forecasting method, the RMSE skill score allows—with appropriate caveats—comparison of forecasts made using different models, across different locations and time periods.

1. Introduction

The power grids, which transmit and distribute electricity to end users, are being monitored and controlled by system operators at all times to ensure reliable power delivery. Considering that solar and other renewable energy sources are inherently variable, and that utility-scale energy storage is not economically viable globally yet, operational excellence of the power grids can benefit from accurate solar forecasts.[1] Consequently, reliable and well-characterized solar forecasting tools and methodologies are becoming essential, and are considered of high value (Martinez-Anido et al., 2016; Huang and Thatcher, 2017; Antonanzas et al., 2017; Klingler and Teichtmann, 2017).

Surface shortwave radiation is unavailable at night and, during daytime, fluctuates as a function of the position of the Sun, cloud cover, aerosols, and other weather variables. Solar forecasts are used by utilities for various reasons: switching energy sources, planning backup generators, calculating reserves, and energy trading. The time horizons covered by modern solar forecasting typically range from a few seconds to a few days. Over the last decade, the literature on this topic has bloomed. A wide spectrum of methods, either physics-based (e.g., sky or shadow imagery, remote sensing, or numerical weather prediction), data-driven (e.g., time series, spatio-temporal statistics, or machine learning), or a combination of both (e.g, hybrid models), have been proposed (see Blaga et al., 2019, Yang et al., 2018; van der Meer et al., 2018; Voyant et al., 2017; Antonanzas et al., 2016; Ren et al., 2015; Inman et al., 2013, for reviews). Furthermore, the existing studies span a range of time intervals and locations, with contrasting weather conditions. Because of these differences, the field would benefit from having a general verification framework for forecast analysis, as well as for the standardization of accuracy measures or metrics[2] for forecast comparison.

This article has three missions. The first is to introduce the distribution-oriented forecast verification framework to the solar forecasting community. The idea of using distributions—in particular the joint distribution of forecasts and observations—originates in the work of Murphy and Winkler (1987). A joint distribution contains all time-independent information relevant to verification. As such, it offers a more detailed view than the traditional measure-oriented approach in terms of forecast analysis. The second mission is to recommend an accuracy measure that should be universally reported in deterministic (also known as single-valued or point) solar forecasting studies—the root mean square error (RMSE) skill score[3] based on the optimal convex combination of climatology and persistence. Since the skill score is able to reflect the inherent difficulties in different forecasting situations, it allows forecast comparisons based on relative improvements, rather than on an absolute error size. The third mission is to look into a series of practical issues in terms of forecast verification, such as data processing or implementation of the reference forecasting methods, with the goal of helping users better understand the relative strengths and weaknesses of various forecasting models in a uniform manner.

Even though the present authors represent a broad range of active researchers in the solar forecasting community, there will always be difficulties in gaining universal consensus on the appropriate measures and methods to be used in general, or even more so in various specific cases. Nevertheless, the authors hope that the forecast verification procedure proposed here can lead to a greater interpretability of results, and even in direct—"apples to apples"—comparisons of techniques. An ultimate goal is to establish the best practices that can be upgraded and refined as the body of knowledge and experience grows.

The organization of this study is as follows. The forecast verification problem and the perceived difficulties are elaborated in Section 2. Distribution-oriented forecast verification methods are discussed and exemplified in Section 3. The recommended accuracy measure is justified in Section 4, alongside with some discussions on practical concerns. Section 5 concludes with a series of recommendations.

2. Problem description

Solar forecasting is a term applied to any predictive form of estimating the solar energy resource ahead of time. With a fast-growing global portfolio of solar energy installations using various technologies, the need for solar forecasting to facilitate improved operations and electricity market compatibility is paramount. A rapidly expanding scientific community in the subdomain of energy forecasting has contributed numerous methodologies and approaches towards solar forecasting (Hong et al., 2016). Accuracy is a major goal of most, if not all, forecasters. The variability in solar irradiance intrinsically governs predictability (Pedro and Coimbra, 2015). Therefore, it is particularly interesting to compare forecasts generated by different models, using data from different locations,[4] or different time periods (Yang, 2019a).

Current methods of solar forecast verification are mostly limited to using measures as indicators of goodness of forecasts. In other words, solar forecasters compare the performance of different models based on

[1] Generally, the phrase "solar forecasting" refers to both solar power forecasting and solar irradiance forecasting. For the latter, forecasters are interested in two irradiance quantities, namely, global horizontal irradiance (GHI) and direct normal irradiance (DNI). Methodologically, there is not much difference in treating GHI and DNI. However, DNI forecasts are usually less accurate due to the larger variability in DNI than in GHI.

[2] "Measures" and "metrics" are distinct concepts in *measure theory*. A measure μ on a set X is a mapping $\mu\colon \mathscr{A} \to [0, \infty]$ defined on a σ-algebra $\mathscr{A}$ that satisfies non-negativity, null empty set, and σ-additivity, that is $\mu(A) \geqslant 0 \;\; \forall \;\; A \in \mathscr{A}$, $\mu(\varnothing) = 0$, and $\mu(\sqcup_{j\in\mathbb{N}} A_j) = \sum_{j\in\mathbb{N}} \mu(A_j)$, where symbol $\sqcup$ denotes disjoint union (Schilling, 2017). On the other hand, a metric is a distance measure $d\colon X \times X \to [0, \infty]$ that satisfies definiteness, symmetry, and triangle inequality, that is $d(x, y) = 0$ iff $x = y$, $d(x, y) = d(y, x)$, and $d(x, y) \leqslant d(x, z) + d(z, y)$, $\forall \;\; x, y, z \in X$. (Schilling, 2017). Nonetheless, moving out from measure theory, the two terms are often used interchangeably, e.g., "accuracy measure" and "error metrics" use the words "measure" and "metric" in their everyday sense. To most forecasters, especially forecast practitioners, they both refer to functions of forecast errors, such as mean bias error (MBE), mean absolute error (MAE), or root mean square error (RMSE).

[3] Skill score is also known as *forecast skill*. It is a class of accuracy measures that gauge the relative improvement of a method over a reference method, see Eqs. (3) and (4) below.

[4] Verification of forecasts, particularly those made by a numerical weather prediction (NWP) model, can be carried out spatially (see Gilleland et al., 2010). Spatial averaging and spatial scale have a strong impact on forecast accuracy (Lorenz et al., 2016). This work, however, is constrained to the verification of forecasts at point locations.

some error metrics, and can then draw conclusions. Under this type of verification procedure, any conclusion is ambiguous in at least two ways: (1) it is unclear what the forecast objective is, and (2) it is unclear how the model of interest performs against other models that are not included in the study. These problems are described in Sections 2.1 and 2.2, respectively.

2.1. What is a good forecast?

The word "objective" refers to goals given to a forecaster prior to verification. It is natural to think of the objective as "small RMSE," "high skill score," or "high economic value." Nonetheless, these objectives often lack generality, and can even be conflicting at times. In that, one may end up collecting a large, and possibly redundant, set of error metrics. This is exemplified by the work of Zhang et al. (2015), in which a suite of 17 metrics was assembled based on a lengthy discussion process that involved stakeholders from both the meteorological and power systems communities. In other cases, new metrics are proposed to meet a specific objective. This is exemplified by the work of Vallance et al. (2017), in which the ability to forecast ramps in irradiance transients is gauged by two new metrics.

Assembling or introducing new members to a pool of error metrics is meaningful to the field of solar forecasting. By presenting a wide spectrum of error metrics, forecasters are able to choose freely among the metrics that can "best" highlight the strengths of their results. There are many studies that propose, contrast, and recommend error metrics to forecasters (e.g., Vallance et al., 2017; Zhang et al., 2015; Hoff et al., 2013; Beyer et al., 2009). However, despite the well-argued discussions, these works can rarely change another forecaster's sentiment towards some specific metrics, if they are perceived as having important advantages or disadvantages. Hence, for each argument that favors a metric, one may find a counter-argument against it (see Chai and Draxler, 2014; Willmott and Matsuura, 2005).[5] Furthermore, since there are countless publications that discuss and conclude that one metric is better than the other, it is not difficult to cite those articles that support any choice the author wishes to make (Chai and Draxler, 2014). The obvious consequence is a field with diverse usage of error metrics. Nonetheless, this is not unique to the emerging field of solar forecasting. Historically, the lack of unified forecast verification procedure has been discussed by many experts from other relatively mature fields (e.g., Murphy and Winkler, 1987; Armstrong, 2001; Fildes et al., 2008), but nothing seems to have changed (Gneiting, 2011).

At this stage, it is essential to ask the question: "what is a good forecast?" It is known, *a priori*, that different metrics favor different forecasts. To put this issue in perspective, a simulation study is presented. Suppose diurnal variation of the hourly clear-sky index, i.e., the ratio between the global horizontal irradiance (GHI) and clear-sky GHI, at an arbitrary location follows:

$$\kappa_t = 1 - z_t^2, \tag{1}$$

where κ_t denotes the clear-sky index at time t, $z_t \sim \mathcal{N}(0, \sigma_t^2)$, and σ_t^2 follows:

$$\sigma_t^2 = 0.15 z_{t-1}^2 + 0.3\sigma_{t-1}^2 + 0.07. \tag{2}$$

With initial values $z_0 = 0$ and $\sigma_0^2 = 0.01$, the κ time series is simulated for 55 daylight hours (herein defined to be data points with a zenith angle <85°), or 5 days. The simulated data points are tabulated and plotted in Table 1 and Fig. 1, together with the corresponding clear-sky GHI (c) values. The McClear model (Lefèvre et al., 2013) is used to estimate c.

Based on the simulated time series, three forecasters are asked to generate forecasts. The *novice* has no skill to offer, and thus issues 1-step-ahead persistence forecasts on κ, i.e., $\phi_t = \kappa_{t-1}$, where symbol "ϕ" denotes the forecast clear-sky index. The *optimist* knows it is sunny in that location, and always uses $\phi_t = 0.95$. The *statistician* has knowledge about the inherent model, and thus issues the true conditional mean as forecasts, i.e., $\phi_t = 1 - \sigma_t^2$. These ϕ values are then converted to GHI forecasts with clear-sky GHI values at the forecast timestamps. The results in terms of three error metrics, namely, mean bias error (MBE), mean absolute error (MAE), and RMSE[6] are shown in Table 2. The results are inconclusive, because each forecaster is best in terms of a particular error metric.

The result of the above simulation study contradicts the common belief that knowing the inherent (physical or statistical) process is the determining factor behind making good forecasts. This contradiction is attributed to how the goodness of forecasts is defined. To most solar forecasters, "good forecast" is implicitly equivalent to "small error." However, the pitfalls of this definition become apparent whenever contradicting rankings of models materialize. In order to resolve such contradictions, solutions might be obtained from the field of meteorology, where forecast verification is well studied.

Murphy (1993) outlined three types of goodness that jointly define a good forecast:

1. *consistency*—correspondence between forecasts and judgments;
2. *quality*—correspondence between forecasts and observations; and
3. *value*—incremental benefits of forecasts to users.

2.1.1. Consistency

Consistency is quite an abstract concept: a forecast is consistent if it corresponds with the forecaster's best judgment. Murphy (1993) argued that such a judgment must contain an element of uncertainty, because the forecaster's knowledge on the forecasting task is necessarily incomplete. In probabilistic forecasting, consistency can be ensured by adopting *strictly proper scoring rules* (Gneiting and Raftery, 2007). With that, forecasters are rewarded with the best scores if and only if their forecasts correspond with their judgment (Murphy and Winkler, 1971). The Brier score and continuous ranked probability score (CRPS), both of which frequently used in probabilistic solar forecasting, are both strictly proper (Gneiting and Raftery, 2007).

On the other hand, in deterministic forecasting, forecasters have to translate their probabilistic judgment through a statistical functional,[7] $T(F)$, which summarizes the forecast distribution, F. The reader is referred to Gneiting (2011) for the formal definition. Informally, the scoring function S is consistent if $\mathbb{E}[S(f,x)] \leqslant \mathbb{E}[S(g,x)]$, for all $f \in T(F)$, where f is an evaluation of the functional, g is any forecast, and x is a future observation. This definition implies that S is consistent if and only if any $f \in T(F)$ is an optimal forecast under S. For example, if the mean value of a forecaster's judgmental probability distribution is of interest, then RMSE is a consistent accuracy measure, because RMSE is minimized by forecasting the mean of the predictive distribution. In the above simulation study, the *statistician* provided the optimal forecasts under RMSE. The *optimist*, although winning the competition with

[5] During the initial stage of this study, the original idea was to propose a specific suite of metrics to the community. However, soon it became obvious that it is impossible to make everyone agree. No consensus can be established on things such as whether MAE or RSME should be favored, whether normalized metrics should be used, or with which quantity the RMSE should be normalized: the mean, maximum, 1000 W/m^2, or square root of second moment.

[6] By surveying 1000 recent forecasting papers, Yang et al. (2018) found that there are about 20 commonly used metrics in solar forecasting, with MBE, MAE, and RMSE being the most popular ones. They can thus be considered most typical, which is why they are used in the simulation study. If the forecast and observed GHI at time t are denoted using f_t and x_t, respectively, then MBE $= \sum_{t=1}^{55}(f_t - x_t)$, MAE $= \sum_{t=1}^{55}|f_t - x_t|$, and RMSE $= \sqrt{\sum_{t=1}^{55}(f_t - x_t)^2}$

[7] Any function of a probabilistic distribution is called a statistical functional. Examples of functionals include mean, median, or variance. Generally, it is written as $T(F)$, where F is a distribution.

Table 1
Simulated clear-sky index ($\varkappa$, dimensionless) and clear-sky GHI (c, in W/m^2) data. The $\varkappa$ simulation follows Eqs. (1) and (2).

Day 1			Day 2			Day 3			Day 4			Day 5		
Index	$\varkappa$	c	Index	$\varkappa$	c	Index	$\varkappa$	c	Index	$\varkappa$	c	Index	$\varkappa$	c
1	1.000	169	12	0.987	170	23	0.824	176	34	0.914	182	45	0.691	168
2	0.941	389	13	0.928	389	24	0.749	397	35	0.993	403	46	0.954	380
3	0.997	587	14	0.987	589	25	0.577	597	36	0.955	601	47	0.683	576
4	0.815	739	15	0.919	743	26	1.000	752	37	0.972	753	48	0.989	728
5	0.871	832	16	0.730	836	27	0.561	846	38	0.770	845	49	0.999	820
6	0.999	859	17	0.393	864	28	0.968	874	39	0.912	871	50	0.998	845
7	0.999	817	18	0.873	822	29	0.972	832	40	0.973	827	51	0.895	800
8	0.963	710	19	1.000	716	30	0.952	724	41	0.995	719	52	0.938	687
9	0.995	546	20	0.925	552	31	0.993	558	42	0.989	553	53	0.648	516
10	0.710	340	21	0.984	346	32	0.960	351	43	0.714	345	54	0.960	308
11	0.998	121	22	0.671	124	33	0.992	130	44	0.918	125	55	0.871	105

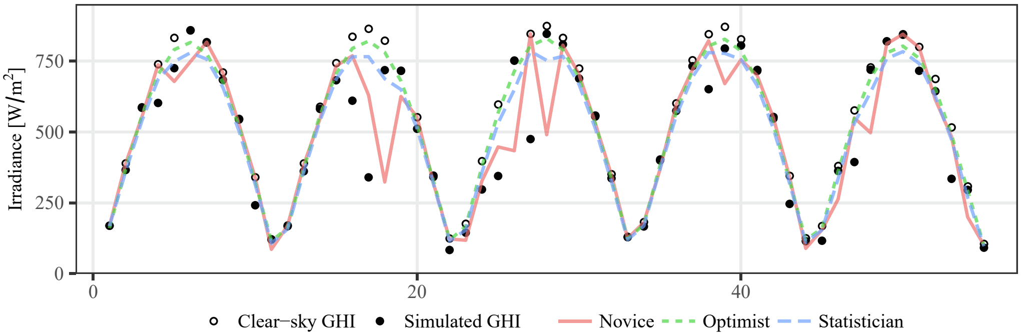

Fig. 1. A window of 55 simulated hourly GHI and clear-sky GHI data points (with zenith angle <85°). Forecasts generated by three forecasters over the same window are overlaid. The *novice* uses 1-step-ahead clear-sky persistence model, the *optimist* always uses 0.95 times of clear-sky GHI as forecasts, and the *statistician* uses the true conditional mean as forecasts.

Table 2
MBE, MAE, and RMSE, in W/m^2, of the three forecasters in the simulation study. Column-wise best results are in bold.

Forecaster	MBE	MAE	RMSE
Novice	**−1.32**	79.80	127.12
Optimist	33.45	**53.96**	100.51
Statistician	11.43	58.08	**93.97**

respect to MAE, did not provide optimal forecasts under MAE (MAE is minimized by forecasting the median of the predictive distribution).

The underlying assumption of using consistency as a measure of goodness of forecasts is that the forecasters receive a *directive* in the form of a statistical functional to transform their probabilistic judgment into a deterministic forecast. For instance, the directive could be "forecast the mean of your probabilistic judgment." Only then, a scoring rule can be identified as consistent if it is optimized by the chosen directive. However, Jolliffe (2008) noted that the very definition of "consistency" is circular: a forecaster can elect to start by choosing a scoring rule. Once the forecasts are made by optimizing the scoring rule, the consistent directive naturally follows.

Consistency implies a theoretical guideline for choosing the most appropriate accuracy measure during forecasting. For most statistical and machine-learning models, the model parameter or weights are estimated or fitted according to some cost function. In this regard, a consistent error measure should be used during verification. For instance, the ordinary least squares regression minimizes the sum of squared errors, hence, RMSE is an appropriate metric to report. Nonetheless, such guideline might not favor forecast comparison in practice, since it would further divide the field. Further studies are necessary.

2.1.2. Quality

Quality is a familiar concept to all solar forecasters, as it refers to the correspondence between forecasts and observations. For example, MAE and RMSE are both measures that assess the overall accuracy of forecasts. Accuracy is an aspect of forecast quality. It can be interpreted through quantitative measures. Besides accuracy, other aspects of forecast quality known to solar forecasters, such as bias, association, skill, or uncertainty, can be assessed through MBE, correlation, skill score, or variance. In forecast verification, the traditional way of comparing measures, may it be positively oriented (the larger, the better, such as skill score), negatively oriented (the smaller, the better, such as RMSE), or center oriented (the closer to a center value, the better, such as MBE), is known as the *measure-oriented approach*.

As mentioned earlier, one disadvantage of using the measure-oriented approach is the subjectivity in choosing measures. Since selecting which measures to report is essentially a decision that is internal to a forecaster, the reasons behind that selection are by default unknown to anyone who is observing the forecast verification procedure from an external view point. In academia, forecasters are authors, whereas observers are editors, reviewers, and readers. If the *optimist* in the simulation study only reports MAE in an article, the observers will not be able to fully realize the underlying pitfalls of those forecasts, but will have no other choice than to accept their results. The simulation study above might over-simplify the state-of-the-art solar forecasting scenarios in comparison with actual forecasting models. Because those are typically much more complex, it would be even more difficult to interpret their results through only a few measures.

A related question is, if two forecasting methods yield the same MBE, RMSE, or skill score, are they equally good? The obvious answer is "No." Measures only provide an *overall* assessment of forecast quality. Since error metrics are often computed based on a collection of samples (e.g., rolling hourly forecasts made over a year), this gives infinite ways to result in the same error-metric value. The reader is referred to Fig. 1

in Vallance et al. (2017) for an example on how two sets of drastically different forecasts can lead to the same RMSE. One solution frequently being used in the solar forecasting literature is to report the regime-dependent error metrics, i.e., to differentiate the errors by classes of prevailing situations. For instance, one can separately report errors for overcast-, clear-, and all-sky conditions. Alternatively, one can also report errors for different times of day, different times of year, or different day types. However, the dimensionality of forecast verification scales with the number of classes, e.g., an RMSE table will become three, if three sky conditions are analyzed separately, or ten, if ten day types are defined. The error contingency table often gets out of control quickly. What has just been discussed is known as *forecast analysis*, which is generally defined as the procedure selected to understand the *composition* of the overall quality.[8]

Since both assessment and analysis of forecast quality are driven by the information embedded in the forecast–observation pairs, it is useful to define the total amount of information available to a forecaster during verification. By defining the entirety of information, a forecaster is no longer limited by the set of summary statistics. Stated differently, if the temporal sequence of forecast–observation pairs is not of interest, the joint distribution of forecast and observation can be used to study the skillfulness of the forecasts, since it contains *all* time-independent information relevant to forecast verification. This *distribution-oriented approach* to forecast verification was proposed by Murphy and Winkler (1987). It has gained high popularity in the field of meteorological forecasting, but is less known by solar forecasters.

This particular framework needs to be discussed because it provides an alternative view to the traditional measure-oriented approach. It offers high flexibility in terms of accessing the information. In fact, the majority of graphical methods (e.g., Taylor diagram, target diagram, or error heat map) and accuracy measures (e.g., MBE, RMSE, or Kolmogorov–Smirnov test integral) known to solar forecasters are specific methods under this general framework. More importantly, the Murphy–Winkler framework is augmented by Bayes' theorem, in that the joint distribution can be written equivalently as the product of marginal and conditional distributions, making the embedded information more accessible. Last but not least, the distribution-oriented approach establishes communication between forecast quality and accuracy measure. Aside from those aspects of forecast quality mentioned earlier, other aspects such as reliability, resolution, or discrimination can easily be defined and quantified. Whereas more details on the Murphy–Winkler framework are provided in Section 3, with a case study, it is noted that the framework is essential to understanding the goodness of forecasts.

2.1.3. Value

Value relates to the benefits realized, or cost incurred, by individuals or organizations who use the forecasts during their decision making. Murphy (1993) pointed out that the forecasts by themselves possess no intrinsic value, because they only acquire value through influencing the decisions made by their users. Most often, the value of solar forecasts is translated into and measured in monetary units.[9] For instance, by reducing the RMSE of the forecasts by x W/m^2, the owner of a photovoltaic (PV) system with energy storage may expect to gain an additional \$$y$ per year through optimizing the feed-in strategy of the system. In a detailed case study, Law et al. (2016) discussed the benefits of improvements in irradiance forecasting for a concentrated solar thermal power plant in this context. An alternative view was given by Antonanzas et al. (2017), where they compared the profit from different forecasting methods with respect to that from perfect forecasts.

Naturally, such benefits or costs depend on the characteristics of a particular decision-making problem. Thus, the third type of goodness is not under the control of forecasters, but is determined and appreciated by decision makers. Furthermore, this goodness of forecast is non-transferable by default. That is, one cannot simply scale the value realized by others, using the characteristics of the problem at hand. Because of the different courses of action and payoff structures available to different decision makers, there is little reason to assume an *ex post* value would apply in an *ex ante* study. A good forecasting strategy that creates high value to some users might not be appreciated by others.

In parallel, it is believed that for a fixed and well-defined decision-making problem, the mapping between quality and value is *monotone*. In other words, higher forecast quality expectedly corresponds to higher value. In principle, this gives the forecaster the necessary motivation to provide the best possible (and hopefully "optimal") forecasts.

To give a perspective on what a "well-defined decision-making problem" can actually be, the case of the Australian National Electricity Market (NEM) is considered. In NEM, conventional generators submit bids every five minutes to the Australian Energy Market Operator. The latter tries to see how far up the bid stacks they have to proceed to meet their forecast net load (regional load forecast minus forecast of domestic and commercial PV generation). If the conventional generators miss their promised amount by more than a given tolerance, either above or below, they are penalized. There is presently a dramatic expansion in solar farm construction and these installations are price takers, i.e., not involved in making the spot market price. Hence, the solar plants would not be fined for poor forecasts, but could be curtailed whenever necessary. Under this regime, the decision-making problem might not be well defined, because the plant owners could always use the highest possible power generation as forecasts, since there is no monetary penalty imposed on over-forecasts. To level the playing field in a new regime, the solar plants could be penalized too, if they do not meet their forecasts. Given the new payoff structure, the cost of over-forecasts can be set to be equivalent to the cost of running spinning reserve to fulfill the difference between the forecast and the generated solar electricity. Similarly, the cost of under-forecasts would be the lost revenue that would have been generated if that extra potential energy were sold at the prevailing spot price at the time. In this case, the decision-making problem is well defined *if* the two costs are the same, then encouraging the forecasters to submit their optimized forecasts truthfully.

2.2. The skill score

The three types of goodness defined by Murphy (1993) provide a clear objective during forecast verification—while maintaining consistency, one should aim at maximizing quality. However, having a well-defined objective only helps the forecasters to analyze and thus make conclusions based on their own forecasting experiment. As the specialized literature expands rapidly, it is unrealistic to expand the scope of the experiment simultaneously, that is, to include all previously proposed methods as benchmarks. The obvious reason is that the data (information) available to one forecaster might not be available to others. Similarly, not all types of information available at one location or time is available at other places or times. Thus, to guarantee progress in the field, the community is forced to make comparison among different research works, based on a variety of reported measures of forecast quality.

2.2.1. A false sense of cross-scenario comparability

The variability of solar irradiance depends on climatic features, geographical location, timescale, and time period, among other factors. Even if the same forecast-generating strategy is employed, the hourly

[8] A related issue is how to decompose these overall quality metrics. Some options are discussed in Section 3.

[9] Other measures of value could be in terms of stability of the grid, customer satisfaction, etc. But how the value is measured does not affect the discussion here.

forecasts made for a location with predominant clear-sky conditions will have a significantly smaller RMSE as compared to 10-min forecasts made for a site with a tropical climate, where cloud formation is rapid and difficult to predict. Hence, if one wishes to compare forecast skills, some form of scaling (normalization) is needed for scale-dependent errors, such as MAE or RMSE. In a recent review paper, Blaga et al. (2019) used the normalized RMSE (nRMSE) as a basis for such comparisons.

The particular form of normalization considered in Blaga et al. (2019) is through the mean of the observations, i.e., nRMSE is computed by dividing RMSE (e.g., in irradiance unit) with the mean of the observations. Whereas the final conclusion—nRMSE reported in the solar forecasting literature is getting smaller over the years—is factual, the methodology (directly comparing nRMSE results from various publications) used by the authors can be misleading. "Researchers are getting better at forecasting solar" is *a priori* knowledge, and it is easy to find evidence supporting that. However, nRMSE gives a false sense of cross-scenario comparability, which cannot be used to justify that one forecaster has better skill than the other. The word "cross-scenario" refers to forecasts generated for data with different predictability.[10] One example is shown in Fig. 2, where RMSE and nRMSE lead to contradictory conclusions.

As mentioned earlier, forecast error is tightly linked to predictability, which is related to variability and uncertainty. Equivalently, it can be stated that conditions with higher variability and uncertainty are harder to forecast, thus leading to larger expected errors. In forecasting, variability and uncertainty are often quantified by step change and variance, respectively. Since the ultimate aim is to have a measure that quantifies forecast skill, its dependency on variability and uncertainty has to be minimized if not removed completely. It is now clear that mean-normalized nRMSE cannot be used to compare forecast skills in general, because the mean is related to neither variability nor uncertainty. The same arguments can be applied to range-normalized nRMSE, max-normalized nRMSE, capacity-normalized nRMSE, and similar variants of nMAE.

2.2.2. On the propagation of normalized accuracy measures in solar forecasting

Normalized accuracy measures are popular in solar forecasting (Blaga et al., 2019). This is in contrast to the field of meteorology, where normalized accuracy measures are rarely used. For instance, in the book by Jolliffe and Stephenson (2012), there is not a single sentence that discusses normalized accuracy measures. Similarly, no trace of normalized accuracy measures can be found in Hyndman and Koehler (2006), in which the accuracy measures are discussed in the context of general-purpose univariate time series forecasting. Hence, some *possible* explanations on why normalized accuracy measures is so popular in solar forecasting are given next.

The notion of normalization develops naturally when the forecast quantities are at different scales. On this point, the class of accuracy measures based on percentage errors, such as the mean absolute percentage error (MAPE), needs to be discussed. Measures based on percentage errors are not quite feasible in high-resolution solar forecasting since the irradiance and PV-generated power is near zero during early mornings and late afternoons, or when the clouds move in. Although the early morning and late afternoon cases can be trimmed with a zenith-angle filter before verification, a few missed forecasts on large irradiance swings during mid-day are enough to result in a very large MAPE. Therefore, to allow the forecast errors to be interpreted as a realistic percentage, the normalization is taken out of the summation, i.e., normalization is performed after aggregation. No such concerns exist when dealing with longer-term forecasts of global irradiance, e.g., on a daily basis.

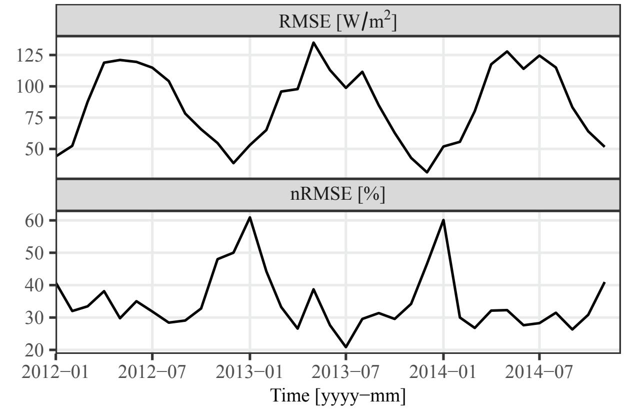

Fig. 2. Monthly averaged day-ahead GHI forecast errors—RMSE [W/m^2] and nRMSE [%] normalized by monthly mean irradiance values—in Hamburg, Germany (53.63°N, 9.99°E), produced by the Integrated Forecasting System (IFS), an NWP system developed by the European Centre for Medium-Range Weather Forecasts (ECMWF). It is evident that winter months have small RMSE but high nRMSE, which might be confusing.

Normalized accuracy measures are used in wind forecasting, a more developed sub-domain of energy forecasting (Hong et al., 2016). In an effort to standardize metric usage in wind forecasting, Madsen et al. (2005) noted that the purpose of using normalized accuracy measures is to produce results independent of wind farm sizes. In addition, the authors recommended normalization by the installed capacity or mean observation. At that time (2005), published studies on solar forecasting were rare. When the field of solar forecasting started to bloom in the early 2010s, such normalization was adopted by solar forecasters, to whom wind forecasting was the most relevant literature to follow.

Another compelling reason why normalized accuracy metrics are frequently used in solar (or wind) forecasting is that the end users (or "stakeholders") are typically electrical engineers, business analysts or financial experts. These professionals are very familiar with percents, but not with a solar radiation unit such as W/m^2 or a power unit such as MW. From this standpoint, the use of normalization is essentially dictated by the necessity for the end users to understand and correctly use the forecast results. Nonetheless, grid operators almost never compare their forecasts across seasons, timescales, and let alone to forecasts of other grid operators. Therefore, in that context, the choice of normalized accuracy metrics is for convenience and internal communications, since a percentage metric is more accessible to a non-technical audience (decision-makers) than a MW metric. Since for grid operators the normalizing quantity, i.e. the denominator, is either constant (peak load) or similar (average load), the normalization does not affect the ranking of forecast accuracy. The grid operator preference for normalized accuracy metrics should *not* be construed as a decisive motivation to report only normalized error metrics in the academic community, because the global inter-comparability of forecast quality needs to be maintained.

2.2.3. Problems with the skill score

Since normalized accuracy measures cannot be used to compare forecasts made at different locations and timescales, an alternative has to be sought. In modern solar forecasting, one of the first attempts to formally address the problem of comparability was made in the early 2010s by Marquez and Coimbra in several publications (Marquez and Coimbra, 2011; Marquez and Coimbra, 2013; Coimbra et al., 2013).[11] In those studies, a well-known concept in meteorological forecasting

[10] If two sets of forecasts are generated for datasets with the same variability, e.g., two PV systems with different nominal power at the same location, nRMSE is a good normalization choice.

[11] Another early discussion was given by Beyer et al. (2009).

called *skill score* was introduced to the young field of solar forecasting. In the field of meteorology, the skill score, s, can be defined based on some measure of accuracy A, namely,

$$s = \frac{A_f - A_r}{A_p - A_r}, \tag{3}$$

where A_f, A_p, and A_r are the accuracy of the forecasts of interest, accuracy of the perfect forecasts,[12] and accuracy of the reference forecasts, respectively (Murphy, 1988). For instance, s based on RMSE is

$$s = 1 - \frac{\text{RMSE}(f, x)}{\text{RMSE}(r, x)}, \tag{4}$$

where f, r, and x are forecasts of interest, reference forecasts, and observations, respectively. For N samples,

$$\text{RMSE}(f, x) = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(f_t - x_t)^2}, \tag{5}$$

$$\text{RMSE}(r, x) = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(r_t - x_t)^2}. \tag{6}$$

The skill score s is often expressed as a percentage, representing the percent improvement in accuracy of the forecasts over the forecasts produced by a reference method. If $s > 0$, the forecasts of interest have a smaller RMSE than that of the reference forecasts, otherwise, $s \le 0$ indicates that the model of interest fails to outperform the reference forecasts. There are, however, two problems with using s to compare forecasts: (1) the choice of accuracy measure can be arbitrary, and (2) the choice of the reference forecasting method can be arbitrary.

The first problem can be understood with a simple example. The computation of s requires a measure of forecast accuracy, A, which is based on a scoring function. Depending on the choice of A, s can be quite different. For instance, suppose $\text{RMSE}(r, x) = 200$ W/m^2 and $\text{RMSE}(f, x) = 100$ W/m^2, then $s_{\text{RMSE}} = 0.5$. However, when the mean square error (MSE) is used, $s_{\text{MSE}} = 1 - (1 - s_{\text{RMSE}})^2$ is boosted to 0.75. Whereas the conversion between s_{RMSE} and s_{MSE} is straightforward, s calculated based on other metrics, such as MAE, would be different, and cannot be inferred from s_{RMSE} or s_{MSE}. Hence, there is no obvious solution to this but to recommend a consensus. At the moment, RMSE is the most common form of A in the literature (Blaga et al., 2019; Yang et al., 2018), and thus should be predominantly used in skill score computation. (This choice is discussed further in Section 4.1.) Hereafter, the symbol s only denotes s_{RMSE}, unless otherwise stated.

One remedy to the second problem is to use a universally-accepted naïve reference model, so that s can be used—with appropriate caveats—to compare the accuracy of forecasts made across different locations or time periods. Skill score is built upon the notion that the forecast errors of a "no skill" reference method should sufficiently reflect the difficulty of the forecasting scenario. In business forecasting, random walk is often used as the reference, and the relative performance of the model of interest is gauged using the Theil's U statistic, a concept similar to skill score (Makridakis et al., 2008). In meteorology, the so-called "climatology" is often used as the naïve reference (Jolliffe and Stephenson, 2012). In deterministic solar forecasting, one of the most popular naïve reference methods—for intra-hour and intra-day forecasts—is the clear-sky adjusted persistence, or simply, *clear-sky persistence*. Clear-sky persistence is conceptually no different from the seasonal naïve method described in Makridakis et al. (2008). That is, persistence forecasts are made based on the clear-sky index, and then adjusted back to irradiance using the clear-sky irradiance at each forecast timestamp. The clear-sky model used in clear-sky persistence can effectively describe two seasonal cycles (a yearly cycle and a daily cycle) in an irradiance time series. Some alternative seasonality adjustment approaches are discussed in Appendix A. As for day-ahead solar forecasting, the *24-h persistence*[13] and the *daily-average climatology* are popular. In this paper, an additional naïve reference, namely, the optimal convex combination of climatology and persistence is introduced, which is guaranteed to outperform both individual reference methods. Such recommendation is addressed at length in Section 4.2.

2.3. The skill score defined by Marquez and Coimbra (2011)

The skill score is not limited to the verification of deterministic forecasts of continuous random variable. It is also used in verification of deterministic forecasts of binary events (e.g, Gilbert skill score or Doolittle skill score), multi-category events (e.g., Gandin and Murphy score), and probabilistic forecast verification (e.g., Brier skill score or CRPS skill score), as described by Jolliffe and Stephenson (2012). While the reader is referred to Jolliffe and Stephenson (2012) for more details on the skill score concept, the version proposed by Marquez and Coimbra (2011) needs to be discussed. In deterministic solar forecasting, their version is one of the most notable alternatives to the skill score defined in Eq. (4).

Marquez and Coimbra (2011) proposed their skill score, denoted here as s^*, based on the concept of "uncertainty" (U) and "variability" (V):

$$s^* = 1 - \frac{U}{V}, \tag{7}$$

where

$$U = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(\frac{f_t - x_t}{c_t}\right)^2}, \tag{8}$$

$$V = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(\frac{x_t}{c_t} - \frac{x_{t-1}}{c_{t-1}}\right)^2}, \tag{9}$$

and f, x, and c are forecast, observation, and clear-sky expectation, respectively. Marquez and Coimbra noted that the ratio between U and V can be approximated by the ratio of the RMSE of the model of interest and the RMSE of clear-sky persistence, i.e., $s^* \approx 1 - \text{RMSE}(f, x)/\text{RMSE}(r, x) = s$. However, no detailed theoretical support was given in the different versions of that proposal (Marquez and Coimbra, 2011; Marquez and Coimbra, 2013; Coimbra et al., 2013). Instead, the approximation was demonstrated empirically, using results from several time series models.

By comparing s^* to s defined in Eq. (4), it is not apparent why the approximation between the two skill scores should hold. In fact, s is the RMSE skill score of irradiance (or solar power) forecasts, whereas s^* is the RMSE skill score of clear-sky index forecasts. Stated differently, the two scores verify different forecast quantities—the former verifies irradiance forecasts, and the latter verifies clear-sky index forecasts. However, upon careful analysis, one can arrive at this conclusion: if the κ forecast error and κ reference forecast error are both independent of the clear-sky expectation, then $s^* = s$, see Appendix B. That said, such independence assumption is almost always violated to a certain degree, due to the imperfect clear-sky model, high sensitivity to three-dimensional effects for broken clouds situations, and the high sensitivity to errors in the atmospheric turbidity. Therefore, the clear-sky index

[12] It should be noted that Eq. (4) assumes the RMSE of the perfect forecast accuracy, A_p, is 0. However, in almost all statistical forecasting frameworks, the models would assume some unpredictable white noise, i.e., non-zero RMSE even if a model perfectly describes the data-generating process. Hence, the assumption here is that the $A_p \ll A_r$, so that it can be neglected.

[13] The term "24-h persistence" means the forecast at hour t on the day of interest is the same as the observation made at hour t on the previous day. If the typical operational forecast submission lead time is considered, the method rather needs to be referred to as "48-h persistence," since the observations made two days prior to the operating day will be used as forecasts.

forecast errors are often larger for low solar elevations (Sengupta et al., 2015; Yang, 2020). Consequently, one only observes $s^* \approx s$ in most cases.

2.4. Section summary

This section has discussed and put forward a few new concepts related to deterministic solar forecast verification. In general, there are two approaches for forecast verification, namely, the measure-oriented approach and the distribution-oriented approach. The goodness of forecasts contains three elements: (1) consistency, (2) quality, and (3) value. Particularly interesting is that the distribution-oriented approach can help forecasters assess the quality of their forecasts in a systematic way—by relating various aspects of forecast quality to different accuracy measures—so that they can be interpreted. This is discussed further in Section 3.

Whereas the distribution-oriented approach is primarily recommended for forecast analysis, i.e., to be used within a forecasting case study, the skill score is recommended for cross-work forecast comparison. It is important to note that the measure-oriented and distribution-oriented approaches are complementary and not substitutive. Skill score is computed based on the accuracy measure of a reference model that can sufficiently describe the difficulty (variability and uncertainty) of a forecast situation. It gauges the overall skillfulness of the forecaster. Considering the importance and economic consequences of inter-comparing a variety of forecasting models in practice, the RMSE skill score should be mandated and standardized. It should be noted that whenever climatology, persistence, or their combination is used in a solar forecasting context, it is intended to be applied to the clear-sky index, or other detrended variable of interest. However, the reference forecasts, i.e., r_t in Eq. (4), should be the original variable of interest, i.e., GHI, DNI, or PV power output. This is discussed further in Section 4.

3. Distribution-oriented approach for forecast verification

The distribution-oriented forecast verification framework is quite general. Before one starts to wonder what joint distribution[14] has to do with forecast verification, most likely one has already used this framework. It is common to use a forecast–observation scatter plot to check forecast quality. One may draw some conclusions based on whether the point cloud is centered on the identity line, or how dispersed the scatter points are. In other cases, a forecaster may wish to check how the scatter points are distributed along the x-axis, or whether the spread of forecasts vary for different observation ranges. In fact, most forecast accuracy quantification—visually or through accuracy measures—are just summaries of the joint distributions, or equivalently, the marginal and conditional distributions. The relationship between joint, marginal, and conditional distributions of two random variables can be expressed using Bayes' theorem. When these variables are the forecast and the observations, the same relationship is referred to as Murphy–Winkler factorization in meteorology (Murphy and Winkler, 1987).

Murphy–Winkler factorizations are:

$$p(f,x) = p(x|f)p(f), \tag{10}$$

$$p(f,x) = p(f|x)p(x), \tag{11}$$

where p denotes distribution, f and x represent forecasts and observations, respectively. Eq. (10) is called the *calibration–refinement factorization*, whereas Eq. (11) is called the *likelihood–base rate factorization*. The naming convention is quite intuitive. For example, the $p(x|f)$ term in Eq. (10) describes the spread of the observations, given a particular forecast. For a good correspondence, the forecast is said to be *calibrated* or *reliable*. Mathematically, the forecasts are perfectly calibrated if $\mathbb{E}(x|f) = \int xp(x|f)dx = f$. The reader is referred to Table 3 for an interpretation of other conditional and marginal distributions, and to Murphy (1997) for a list of aspects of forecast quality.

Verifying the above conditional and marginal distributions is equivalent to verifying the joint distribution. For instance, given two sets of forecasts, f_1 and f_2, by comparing $p(x|f_1)$ and $p(x|f_2)$, one can conclude whether one set of forecasts is more reliable than the other; see Moskaitis (2008), Murphy et al. (1989) for case studies. Whereas linking the forecast distributions to aspects of forecast quality provides forecasters with insights regarding their forecasts, such results are easier to interpret if the different aspects of forecast quality can be quantified using measures. For instance, consider the bias–variance decomposition of MSE:

$$\begin{aligned}\text{MSE} &= \iint (f-x)^2 p(f,x)dfdx\\ &= \mathbb{E}[(f-x)^2\\ &= \mathbb{V}(f-x) + [\mathbb{E}(f) - \mathbb{E}(x)]^2\\ &= \overbrace{\mathbb{V}(f) + \mathbb{V}(x)}^{\text{marginal dist.}} - \overbrace{2\text{cov}(f,x)}^{\text{association}} + \overbrace{[\mathbb{E}(f) - \mathbb{E}(x)]^2}^{\text{unconditional bias}},\end{aligned} \tag{12}$$

where the overhead braces show the representation of each term. In this decomposition, $\mathbb{V}(f)$ and $\mathbb{V}(x)$ are variances of forecasts and observations, respectively. Their values can be used as a proxy for measuring the similarity between $p(f)$ and $p(x)$. If the forecasts were perfect, the two marginal distributions would be exactly the same, and so would the variances.[15] Similarly, the $\text{cov}(f,x)$ term can be written as correlation, namely, $\sqrt{\mathbb{V}(f)\mathbb{V}(x)}\cdot\text{cor}(f,x)$, which denotes the *association* between forecasts and observations. Lastly, the $[\mathbb{E}(f) - \mathbb{E}(x)]^2$ term represents the squared unconditional bias, i.e., MBE^2. This example illustrates the complementarity between the measure-oriented approach (e.g., verification using MBE, correlation, or variance of the forecasts) and the distribution-oriented approach (analyzing the joint, conditional, and marginal distributions of forecasts and observations). Some of the widely used measures in solar forecasting, such as MBE, RMSE, or the Kolmogrov–Smirnov test integral, are related to their respective distributions in Appendix C.

Besides the bias–variance decomposition, MSE can also be decomposed following the calibration–refinement and likelihood–base rate factorizations:

$$\text{MSE} = \mathbb{V}(x) + \overbrace{\mathbb{E}_f[f - \mathbb{E}(x|f)]^2}^{\text{type 1 conditional bias}} - \overbrace{\mathbb{E}_f[\mathbb{E}(x|f) - \mathbb{E}(x)]^2}^{\text{resolution}}, \tag{13}$$

$$\text{MSE} = \mathbb{V}(f) + \overbrace{\mathbb{E}_x[x - \mathbb{E}(f|x)]^2}^{\text{type 2 conditional bias}} - \overbrace{\mathbb{E}_x[\mathbb{E}(f|x) - \mathbb{E}(f)]^2}^{\text{discrimination}}. \tag{14}$$

A derivation of these decompositions is shown in Moskaitis (2008). As annotated above the equations, different terms in the decomposed forms explain different aspects of forecast quality.

The *type-1 conditional bias*, $\mathbb{E}_f[f - \mathbb{E}(x|f)]^2$, indicates the degree of correspondence between the mean observation given a particular forecast and the conditioning forecast, i.e., calibration or reliability. Recall that perfect calibration is when $\mathbb{E}(x|f) = f$, so the smaller this term is, the better. *Resolution* accounts for the difference between conditional and unconditional mean observation, which is reflected by $\mathbb{E}_f[\mathbb{E}(x|f) - \mathbb{E}(x)]^2$. If $\mathbb{E}(x|f) = \mathbb{E}(x)$, it means the data samples have no resolution. Since it is desirable to have the generated forecasts to be followed by different observations (so that the forecasts are meaningful), this term should be maximized. This is also reflected by the negative sign in front of that term. The *type-2 conditional bias*,

[14] Formally, we call a function $p(x,y)$ the joint distribution of random variables X and Y if $p(x,y) \geqslant 0$, $\forall\ (x,y)$; $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x,y)dxdy = 1$; and for any set $\mathscr{A} \subset \mathbb{R} \times \mathbb{R}$, $\mathbb{P}[(X,Y) \in \mathscr{A}] = \iint_{\mathscr{A}} p(x,y)dxdy$, $\mathbb{P}[(X,Y) \in \mathscr{A}]$ denotes the probability of (X,Y) in set $\mathscr{A}$ (Wasserman, 2013).

[15] However, having identical variances does not imply identical distributions; and having identical distributions does not imply the forecasts are perfect.

Table 3
Definition, interpretation, and quantification of Murphy–Winkler factorizations (Jolliffe and Stephenson, 2012; Murphy and Winkler, 1987).

Distribution	Definition	Interpretation	(some specific methods of) quantification
$p(x\|f)$	(related to) *calibration*	A set of deterministic forecasts is perfectly calibrated if $\mathbb{E}(x\|f) = \int xp(x\|f)dx = f$.	(1) *calibration*, *reliability*, or *type-1 conditional bias*: $\mathbb{E}_f[f - \mathbb{E}(x\|f)]^2$; (2) *resolution*: $\mathbb{E}_f[\mathbb{E}(x\|f) - \mathbb{E}(x)]^2$.
$p(f)$	(related to) *refinement*, or *sharpness*	Refinement or sharpness is an aspect that usually applies only to probabilistic forecasts (Murphy, 1997). In deterministic forecast verification, if a forecaster produces the same forecast all the time, it is said to be completely unrefined. However, the complete refinement is difficult to define for deterministic forecasting (Murphy et al., 1989), but $p(f)$ has to be equal to $p(x)$ for perfect forecasts.	(1) *Kolmogorov–Smirnov test statistic*: $\max\|F(f) - F(x)\|$; (2) *earth mover's distance* or *first Wasserstein distance*: the area between the two ECDFs. (The formal definition is technical and thus omitted.)
$p(f\|x)$	*likelihood*	If $p(f\|x)$ is zero for all values x but one, the forecast is perfectly discriminatory. If $p(f\|x)$ is the same for all values of x, the forecast is not at all discriminatory.	(1) *discrimination 1*, or *type-2 conditional bias*: $\mathbb{E}_x[x - \mathbb{E}(f\|x)]^2$; (2) *discrimination 2*, or simply *discrimination*: $\mathbb{E}_x[\mathbb{E}(f\|x) - \mathbb{E}(f)]^2$.
$p(x)$	*uncertainty*, or *base rate*	If $p(x)$ is a fairly peaked distribution, the scenario has relatively small uncertainty (and thus easier to forecast) as compared to a scenario where $p(x)$ is fairly uniform.	(1) *variance*: $\mathbb{V}(x)$; (2) *kurtosis*: $\frac{\mathbb{E}[(x-\mathbb{E}(x))^4]}{(\mathbb{E}[(x-\mathbb{E}(x))^2])^2}$.

$\mathbb{E}_x[x - \mathbb{E}(f|x)]^2$, indicates the degree of correspondence between the mean forecast given a particular observation and the observation. Naturally, this term should be as small as possible. Lastly, *discrimination* denotes the difference between the conditional and unconditional mean forecasts, i.e., $\mathbb{E}_x[\mathbb{E}(f|x) - \mathbb{E}(f)]^2$, which indicates how forecasts are differentiated for different observation values. This terms needs to be maximized.

The numerical evaluation of these decomposed factors can be difficult. When Murphy and Winkler (1987) proposed these decompositions, a binary x was used in their case study, which greatly simplifies the computation. In Moskaitis (2008), the evaluation was performed by discretizing the continuous random variable—tropical cyclone intensity—into bins. Recently, Yang and Perez (2019) used kernel conditional density estimation (KCDE) to estimate the conditional expectations, namely, $\mathbb{E}(x|f)$ and $\mathbb{E}(f|x)$, which removes the dependency on binning. The code for the KCDE-based approach is available in the supplementary material of that paper.

In contrast to numerical evaluation of Eqs. (13) and (14), visual inspection is more straightforward and enables a forecaster to appreciate the properties of the forecasts in great detail. In general, visualizing the error distribution is a powerful way of communicating the performance of a model. In line with the Murphy–Winkler factorizations, an x–f scatter plot displays the joint distribution between observations and forecasts, and allows for visualizing the marginal distributions as well as specific conditional distributions.

To exemplify the forecast verification procedure discussed in this section, a case study is presented. Fig. 3 shows the joint and marginal distributions of 24-h-ahead hourly forecasts of global horizontal irradiance (GHI) produced by the North American Mesoscale (NAM) forecast system against the observations collected by the Surface Radiation Budget Network (SURFRAD), at two locations with distinct climate over a period of two years. Whereas the joint distribution at the Desert Rock (DRA) station has approximately equal probabilities on both sides of the diagonal, the forecasts at the Penn. State Univ. (PSU) station overpredict GHI, i.e., the probability density is higher above the diagonal, where $f > x$. A closer examination of the 2D kernel density contours reveals that the NAM forecasts at DRA drift slightly below the identity line for high-irradiance conditions. For mid- and low-irradiance conditions at DRA, the forecasts are slightly above the identity line. This observation warrants an irradiance-condition-based post-processing treatment. Similar observations can be made for forecasts at PSU.

The histograms shown in Fig. 3 denote marginal distributions, $p(f)$ (to the right) and $p(x)$ (on the top). Since the shape of the histograms depends largely on bin width, different choices may affect the forecaster's judgment differently. In this regard, overlaying the empirical cumulative distribution functions (ECDFs) of f and x could be useful at

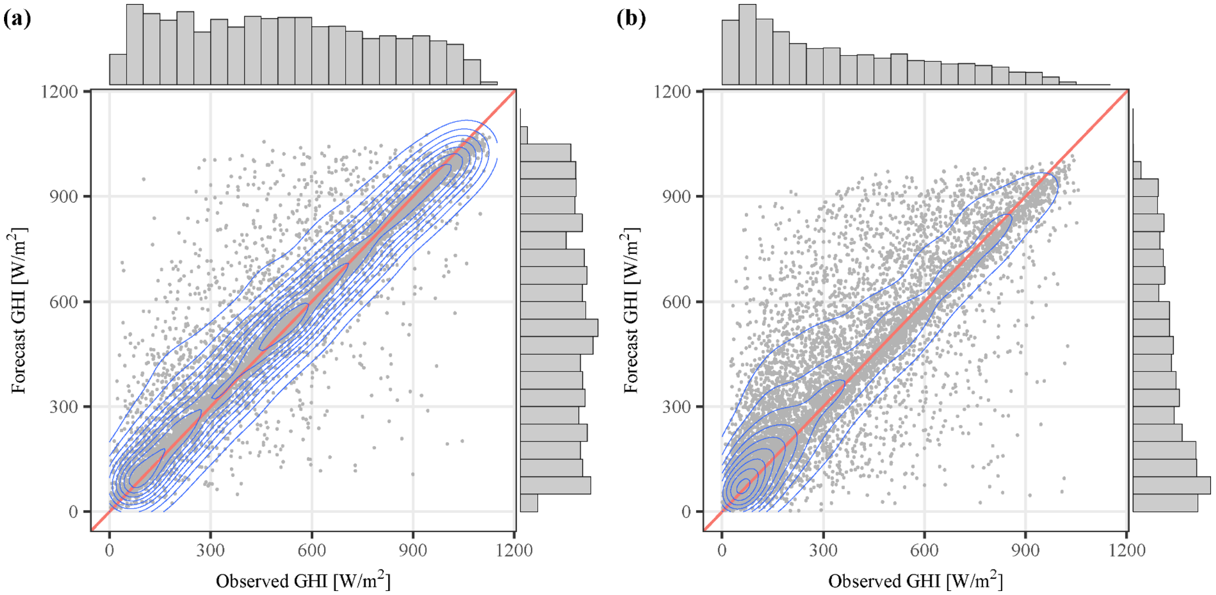

Fig. 3. Joint and marginal distributions of 24-h-ahead hourly NAM forecasts and SURFRAD observations at (a) Desert Rock, Nevada (36.624°N, 116.019°W), and (b) Penn. State Univ., Pennsylvania (40.720°N, 77.931°W), from 2015 to 2016. The contour lines show the 2D kernel densities.

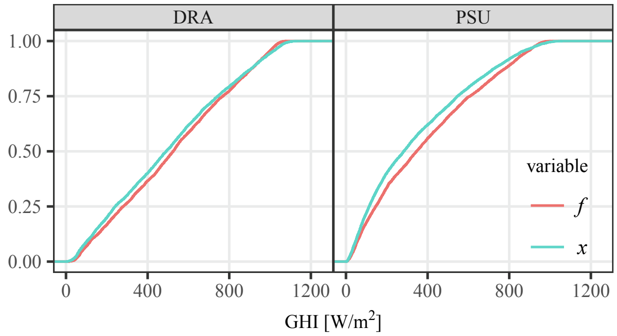

Fig. 4. Marginal distributions of forecasts and observations described in Fig. 3.

times. Fig. 4 demonstrates such plots using the same data. Visually, the ECDFs of forecasts and observations at DRA align better than those at PSU. At PSU, f is stochastically greater than x (the ECDF of f lies below and hence to the right of that for x). Formally, the Kolmogorov–Smirnov test computes the statistic $D_n = \max|F_n(f) - F_n(x)|$, i.e., the maximum absolute distance between the ECDFs of forecasts and observations. In the present case, Kolmogorov–Smirnov tests conducted at the two stations both reject the null hypothesis—the two distributions are equal—at a significance level of 0.05.

As compared to joint and marginal distributions, the visualization of conditional distributions is more challenging. Whereas some authors plot the individual quantiles or use box plots to represent the distributions, ridgeline plots are employed here, see Fig. 5. In this plot, $p(x|f)$ and $p(f|x)$ at both stations are represented using overlapping lines, which create the impression of a mountain range. Fig. 5 (a) and (c) reveal that $p(x|f)$ is mostly centered on the forecast value, i.e., $\mathbb{E}(x|f)$ is close to f, indicating small type-1 conditional bias in NAM forecasts. On the other hand, type-2 conditional bias is found to be significant for high values of x, see $p(f|x)$ for x = 1050 W/m^2 in Fig. 5 (b) and (d).

The distribution-oriented verification technique is often

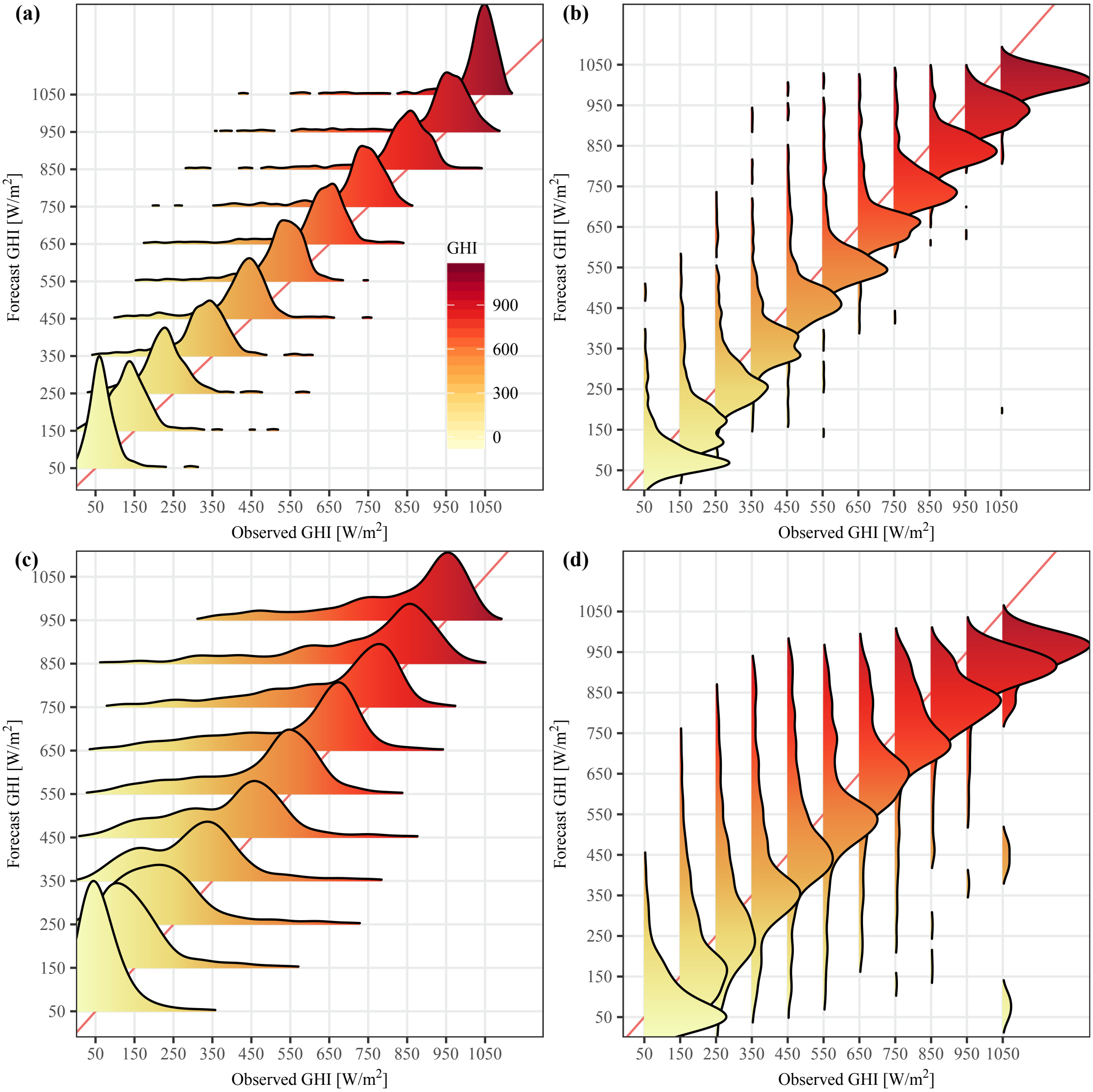

Fig. 5. Conditional distributions of 24-h-ahead hourly NAM forecasts and SURFRAD observations. $p(x|f)$ are shown in (a) and (c) for Desert Rock, Nevada (36.624°N, 116.019°W) and Penn. State Univ., Pennsylvania (40.720°N, 77.931°W), respectively. $p(f|x)$ are shown in (b) and (d) for the two stations, respectively.

Table 4

Bias–variance decomposition (see Eq. 12) and Murphy–Winkler factorization (see, Eqs. 13 and 14) of 1–24-h-ahead NAM (f) against SURFRAD GHI (x), at DRA and PSU stations over 2015–2016. For interpretability, all metrics are written in squared form, so that the bases have the unit of W/m^2, except for correlation ρ, which is dimensionless.

	MSE	$\mathbb{V}(x)$	$\mathbb{V}(f)$	$\rho(f, x)$	$[\mathbb{E}(f) - \mathbb{E}(x)]^2$	$\mathbb{E}_f[f - \hat{\mathbb{E}}(x\|f)]^2$	$\mathbb{E}_f[\hat{\mathbb{E}}(x\|f) - \mathbb{E}(x)]^2$	$\mathbb{E}_x[x_g - \hat{\mathbb{E}}(f\|x)]^2$	$\mathbb{E}_x[\hat{\mathbb{E}}(f\|x) - \mathbb{E}(f)]^2$
DRA	108.29^2	297.70^2	288.42^2	0.94	22.67^2	28.02^2	278.68^2	38.88^2	270.09^2
PSU	155.11^2	269.83^2	275.26^2	0.85	41.73^2	64.41^2	229.86^2	61.24^2	235.40^2

Table 5

Same as Table 4, but tabulating the forecast quality of 24-h-ahead forecasts made using the optimal convex combination of climatology and persistence on clear-sky index.

	MSE	$\mathbb{V}(x)$	$\mathbb{V}(f)$	$\rho(f, x)$	$[\mathbb{E}(f) - \mathbb{E}(x)]^2$	$\mathbb{E}_f[f - \hat{\mathbb{E}}(x\|f)]^2$	$\mathbb{E}_f[\hat{\mathbb{E}}(x\|f) - \mathbb{E}(x)]^2$	$\mathbb{E}_x[x_g - \hat{\mathbb{E}}(f\|x)]^2$	$\mathbb{E}_x[\hat{\mathbb{E}}(f\|x) - \mathbb{E}(f)]^2$
DRA	115.94^2	297.70^2	259.82^2	0.92	6.96^2	22.33^2	274.99^2	60.12^2	240.08^2
PSU	168.06^2	269.83^2	185.12^2	0.79	14.77^2	37.19^2	214.20^2	126.84^2	148.57^2

complemented with *summary measures* of the different aspects of forecast quality. Table 4 shows the quantification of these aspects using the bias–variance decomposition and Murphy–Winkler factorization, as stated in Eqs. (12)–(14). Note that the decomposed terms listed in Table 4 do not add up exactly to MSE, due to the uncertainty introduced during KCDE. Such discrepancy is however small, and thus does not affect this analysis. In terms of correlation, a higher $\rho = 0.94$ is observed at DRA as compared to 0.85 at PSU, indicating a better association between forecasts and observations at DRA. The square of unconditional bias, $[\mathbb{E}(f) - \mathbb{E}(x)]^2$, is also significantly smaller at DRA, agreeing with the earlier observation made using the joint distribution plots. Since a smaller type-1 conditional bias, $\mathbb{E}_f[f - \hat{\mathbb{E}}(x|f)]^2$, means higher calibration—the forecasts at DRA are more reliable than those at PSU. Similarly, smaller type-2 conditional bias, higher resolution, and higher discrimination observed at DRA all lead to the conclusion that the NAM forecasts at DRA have better forecast quality than those at PSU. This conclusion confirms what could be expected because of the much cloudier climate of the latter station.

Based on the case study above, it is evident that the distribution-oriented forecast verification is useful in assisting forecasters to make informed decision based on forecast quality. In other cases, the same methodology can be applied to compare forecasts made using different methods, thus providing more information than using MSE values alone. This verification procedure leads to more meaningful conclusions than statements such as “the MSE at location A is smaller than that at location B, and thus the forecasts at location A are better.” Nevertheless, should one wish to examine the relative accuracy gain from the reference method, the quantification of aspects of forecast quality can be carried out with the reference forecasts, namely, the optimal convex combination of climatology and persistence, as exemplified in Table 5. Comparing the two tables, it is obvious that in the case of NAM, the NWP-based model dominates the 24-h persistence in all aspects except for the unconditional and type-1 conditional bias. Nonetheless, such bias in NWP forecasts can be trimmed with regression-based post-processing, and thus does not affect one’s confidence in opting for the NAM model.

To promote the uptake of this distribution-oriented approach, data and code for reproducing all results appeared in this section are provided in Appendix D.

4. Recommendations and practical concerns

The traditional measure-oriented approach is complemented by the distribution-oriented approach. The two approaches reflect different aspects of forecast quality and help forecasters analyze their forecasts. Generally, forecasters are encouraged to use any meaningful measure to gauge their forecasts. Ultimately, however, a “one-number summary” of forecasts is still highly desirable, especially when scientists bring their forecasts to non-technical personnel, e.g., sales persons, politicians, or the general public. Therefore, in this section, some recommendations and practical concerns regarding the use of skill scores are discussed.

4.1. MBE-, MAE-, or RMSE-based skill score?

The selection of RMSE to evaluate the skill score was made in Section 2.2.3. The arguments for that choice are presented in this section. The skill score belongs to the class of relative measures (Hyndman and Koehler, 2006),[16] which means that a scale-dependent measure is needed for its computation. Since MBE, MAE, and RMSE are currently the most popular metrics (Yang et al., 2018), the discussion below considers all these options.

MBE is defined as $\mathbb{E}(f - x)$, or equivalently, $\mathbb{E}(f) - \mathbb{E}(x)$. An opposite definition, $\mathbb{E}(x - f)$, also exists in the statistics literature, which can be confusing. The latter definition originates from how a predictive model is constructed (see Makridakis et al., 2008).[17] Defining the MBE to be “forecast minus observation” is more natural for solar forecasting, since an over-prediction (where forecasts are, on average, higher than observations) corresponds to a positive MBE, and an under-prediction corresponds to a negative MBE. MBE describes unconditional bias, and most statistical forecasting methods have MBE → 0. State-of-the-art operational solar forecasts would have some form of bias correction implemented, e.g., model output statistics (MOS). Therefore, having small MBE is more of a baseline requirement, rather than a creditworthy feature among state-of-the-art forecasts. Furthermore, the MBE of the reference forecasts from clear-sky persistence has an expectation of zero, and thus makes the skill score undefined. To that end, MBE is unsuitable for skill score computation.

The main difference between MAE, defined as $\mathbb{E}(|f - x|)$, and RMSE, defined as $\sqrt{\mathbb{E}[(f - x)^2]}$, is that the latter penalizes large errors while the former gives the same weight to all errors. Since large errors are particularly concerning for grid integration of solar power (e.g., a loss of load becomes more likely), RMSE is more suitable when a set of forecasts contain several large errors, which is usually the case for solar forecasts. From that standpoint, the percentage improvement in RMSE, in the form of s, might attract more interests than the MAE skill score.

[16] One should distinguish relative measure from measure of relative error. The former performs division after the primary measure is computed, i.e., $\mathbb{E}[S(f, x)]/\mathbb{E}[S(r, x)]$, whereas the latter performs averaging on relative errors, i.e., $\mathbb{E}[S(f, x)/S(r, x)]$, where S is a scoring function.

[17] For a regression model, $y = g(x) + e$, where y is the response, x is the regressor, and bias e is in fact $\mathbb{E}[y - g(x)]$, or observation minus prediction.

The second reason for using the RMSE skill score is related to the distribution-oriented forecast verification. The Murphy–Winkler MSE factorization has been recommended for forecast analysis in the previous section. As a result, RMSE values become readily available after the various aspects of forecast quality are quantified. It must be highlighted that in the field of meteorology, and many other fields such as statistics, researchers generally do not distinguish "RMSE" and "MSE" in their writing (Jolliffe and Stephenson, 2012), simply because RMSE and MSE differ only by a square root. Nonetheless, they should not be mixed up during skill score computations, as evidenced by the example given in Section 2.2.3.

Lastly, the popularity of RMSE is higher than that of MAE, not only in solar forecasting, but in other forecasting domains as well. Gneiting (2011) found that the usage of RMSE in four related domains, namely, forecasting, statistics, econometrics, and meteorology, dominates as compared to MAE. Whereas the precise reasons are unknown, it is hypothesized that consistency might be one of the main reasons, since there are more models minimizing MSE than minimizing MAE. Additionally, squares are more amenable than absolute values in many mathematical operation (Chai and Draxler, 2014). Hence, for this and the above reasons, the RMSE skill score is recommended in deterministic solar forecasting.

4.2. Climatology, persistence, or their convex combination?

There are different definitions of what constitutes a climatology method. Murphy (1988) considered single-valued internal climatology, multiple-valued internal climatology, single-valued external climatology, and multiple-valued external climatology—each resulting in a different skill score expression. Whereas "single-valued" refers to methods that issue a single forecast for all occasions, "multiple-valued" methods issue different forecasts based on some conditional variables (e.g., one value for each season). The definitions of "internal" and "external" are based on whether the reference forecasts are derived directly from the n samples being verified, or derived from historical samples that are not included in the n samples. For instance, given an experimental period of one year, the sample mean, $\mathbb{E}(x)$, of all daytime observations in that year, is called the internal single-valued climatology forecast. Under this definition, the MSE of single-valued internal climatology is simply the sample variance, i.e.,

$$\text{MSE}_c = \mathbb{E}[\mathbb{E}(x) - x]^2 = \mathbb{V}(x). \tag{15}$$

It might be however argued that the multiple-valued climatology would be more appropriate when the forecasts of interest are to be evaluated over a time period longer than a month or a season (Murphy, 1988). For instance, such a climatology forecast is typically constructed for each month or each season. In the most extreme cases, the forecasts issued by multiple-valued climatology could be constructed in a rolling manner, with the averaging window equal to the forecast horizon. For example, in a day-ahead scenario of hourly GHI forecasts, one would use the averaged observed clear-sky index from the last available day as the "climatological" forecasts for each hour of the next day. This day-ahead reference method has been used by Beyer et al. (2009), Perez et al. (2013) among others and has been experimentally validated by Yang (2019c). However, due to the many choices in constructing multiple-valued climatology forecasts, recommending the multiple-valued climatology defeats the purpose of standardizing the reference method used in solar forecasting. For this reason, the multiple-valued climatology will not be discussed further in this article.

On the other hand, Murphy (1992) defined the persistence forecast as a forecast based solely on the value of the variable of interest at an initial time. Given a forecast horizon, h, the MSE of h-step-ahead persistence can be written as:

$$\text{MSE}_p = \mathbb{E}(x_{i-h} - x_i)^2 = 2(1 - \gamma_h)\mathbb{V}(x), \tag{16}$$

where γ_h is the lag-h autocorrelation function. By comparing MSE_p to MSE_c, it is obvious that the relative performance of single-valued internal climatology and h-step-ahead persistence depends on γ_h. When $\gamma_h < 0.5$, $\text{MSE}_c < \text{MSE}_p$. When $\gamma_h > 0.5$, $\text{MSE}_c > \text{MSE}_p$. The two MSEs are equal if and only if $\gamma_h = 0.5$.

To ensure that the value of γ_h, which depends on the association among observations at lag h, does not affect a forecaster's choice of reference method, one can consider a convex combination of the two. The combined forecast is $\alpha x_{i-h} + (1-\alpha)\mathbb{E}(x)$. The optimal value of α can be obtained by differentiating the MSE of the climatology–persistence combination method, and equating it to zero, i.e., $d\text{MSE}_{cp}/d\alpha = 0$. It is straightforward to see that the optimal α is simply γ_h, i.e., $\alpha = \gamma_h$. The MSE of the optimal convex combination of single-valued internal climatology and h-step-ahead persistence is thus:

$$\begin{aligned} \text{MSE}_{cp} &= \mathbb{E}[\gamma_h x_{i-h} + (1-\gamma_h)\mathbb{E}(x) - x_i]^2 \\ &= \mathbb{E}\{\gamma_h[x_{i-h} - \mathbb{E}(x)] - [x_i - \mathbb{E}(x)]\}^2 \\ &= \gamma_h^2\mathbb{V}(x) + \mathbb{V}(x) - 2\gamma_h\text{cov}(x_{i-h}, x_i) \\ &= (1-\gamma_h^2)\mathbb{V}(x), \end{aligned} \tag{17}$$

where cov denotes the covariance. By comparing the expression of MSE_{cp} to MSE_c and MSE_p, it is apparent that $\text{MSE}_{cp} = (1-\gamma_h^2)\text{MSE}_c$ and $\text{MSE}_{cp} = [(1+\gamma_h)/2]\text{MSE}_p$. Since $-1 \leqslant \gamma_h \leqslant 1$, one obtains $\text{MSE}_{cp} \leqslant \text{MSE}_c$ and $\text{MSE}_{cp} \leqslant \text{MSE}_p$, with equality only when $\gamma_h = 0$ and $\gamma_h = 1$, respectively. Consequently, the RMSE of the climatology–-persistence combination is always smaller or equal to that of either reference method, that is,

$$\text{RMSE}_{cp} \leqslant \min(\text{RMSE}_c, \text{RMSE}_p). \tag{18}$$

This condition is a powerful support for choosing the climatology–-persistence combination as the standard of reference method in deterministic solar forecasting. This choice has been discussed by Yang (2019b, 2019c), with extensive empirical validation, for intra-hour, intra-day, and day-ahead scenarios.

4.3. Choice of the clear-sky model

Although it should be obvious at this point, it is re-emphasized that the naïve reference methods in solar forecasting always operate on clear-sky index. A clear-sky radiation model is thus needed to convert all-sky irradiance to clear-sky index and back. When developing forecasts using an NWP model, chances are the model can output the ideal clear-sky irradiance at regular time steps. When these direct outputs are not available, one of the many clear-sky radiation models that are available must be selected and run separately. Several extensive reviews on those models that are typically used in solar applications have been published recently (e.g., Sun et al., 2019; Antonanzas-Torres et al., 2019; Ruiz-Arias and Gueymard, 2018b). The best such models are simpler, but also faster, than those used in NWP models, while providing surface irradiance estimates of similar accuracy.

One particular issue is that most of the high-performance clear-sky models require several inputs, such as total column ozone amount, precipitable water, aerosol optical depth, or aerosol single-scattering albedo. Most generally, these quantities are not observed locally and must be sourced from remote-sensing or reanalysis databases. The process of obtaining such data at the proper temporal resolution from the highest-quality sources is not easy and may appear intimidating to non-experts. On the other hand, very simple radiation models that require only a few undemanding input variables usually have too limited performance for serious work (Gueymard, 2012). Therefore, the McClear model (Gschwind et al., 2019; Lefèvre et al., 2013) might be the best choice for solar forecasters, at least if they do not need to delve into the distant past (since the model delivers data only after 2004-01-01). Being a physical model based on radiative transfer, the performance of McClear is among the best (Ruiz-Arias and Gueymard, 2018a; Sun et al., 2019). McClear is conveniently available as a web service[18] that delivers its predictions at five different time scales (1 min, 15 min, hourly, daily, and monthly) for all locations in the world, and for 2004-01-01 up to two days ago.[19] The R package "camsRad" is also freely available, and offers access to McClear through an API. The only limitation of McClear comes from its spatio-temporal resolution constraints, imposed by the reanalysis data it uses as its main inputs: 0.5° in both latitude and longitude on a 3-hourly basis.[20] Hence, over complex terrain (e.g., mountains) or during fast-changing atmospheric conditions (e.g., dust storms), the McClear predictions might significantly differ from the actual situation at any specific site within its defining 0.5° × 0.5°pixel. In such cases, the forecaster would have the burden of appropriately correcting the clear-sky predictions.

4.4. *Some seemingly trivial implementation issues*

Forecast verification must be based on appropriate out-of-sample tests (see Tashman, 2000, for a review on various test designs). There are several seemingly trivial implementation issues such as data trimming, normalization, counting nighttime hours, and data aggregation, that can strongly affect the verification results.

Data trimming refers to quality control (QC) applied to experimental data prior to forecasting and verification. Owing to factors such as measurement uncertainty or irradiance modeling error, experimental data often contain spurious values. There is no universally-accepted QC procedure (Gueymard and Ruiz-Arias, 2016), but recommended QC methods for surface radiation measurements (Long and Shi, 2008; García et al., 2014) and PV power output (Killinger et al., 2017) do exist. Recently, owing to the advances in remote-sensing technology, satellite-derived irradiance products have been shown to help perform QC on irradiance measurements (Urraca et al., 2017; Perez et al., 2017).

As statistical and machine-learning software packages become more powerful, details of implementation are now more opaque to forecasters. Therefore, forecasters should check the output of each step during forecasting, responsibly, to prevent spurious data from entering the final verification stage. To ensure forecast verification is performed with reasonably trimmed data, visual inspection as outlined in the previous section is necessary. It is however noted that data trimming based on forecast error is *not* recommended. One should not remove a forecast–observation pair just because it produces a large error; instead, the cause behind it should be investigated.

In parallel, there is also no well-accepted answer to the question "whether the nighttime hours should be included during validation." Inclusion of nighttime hours would make the overall error smaller, since the forecasts—0 W/m^2—are perfect during those hours. Hence, nighttime data should always be excluded from verification. This could be ensured by using a zenith angle filter of <85°—a provision that also removes low-sun situations during which measurements are less accurate and irradiance is too low to be of significance in solar power applications. Nonetheless, special care is needed when using the filtered data. For instance, the sample average used in single-valued internal climatology should be based on the filtered data, whereas the lag-h autocorrelation should be computed based on the unfiltered data, to preserve the temporal spacing of the original clear-sky index time series.

In state-of-the-art solar forecasting, it is common to have more than one data source involved (e.g., ground measurements, satellite-derived irradiance, reanalysis data, NWP output, or PV output). Even in the case of verifying NWP forecasts only, one often desires to compare them to ground-based measurements. The issue of data aggregation naturally comes into play because the ground-based measurements are usually at a higher temporal resolution (e.g., 1 min) than the NWP output (e.g., hourly). A related issue is caused by occurrences of data gaps in the data stream. This is a relatively frequent issue, caused by either instrument malfunction or detection of incorrect or suspicious data by the QC procedure. Data gaps create a difficulty when attempting to obtain hourly or daily averages. Some gap-filling or statistical methods do exist, however, to limit the resulting uncertainty in the temporal means.

There are three averaging schemes to aggregate a high-temporal-resolution time series into a low-temporal-resolution one, namely, floor, ceiling, and round. A floor aggregation means that the data within a time interval are aggregated to the earliest timestamp in that interval, e.g., 1-min data points between 1:00 to 2:00 are aggregated and stamped as 1:00. Similarly, the ceiling-aggregation scheme stamps the aggregated data with the last timestamp of an interval, and the round-aggregation scheme collapses the data to the center timestamp of an interval. Experimental irradiance data may follow any one of these conventions (Polo et al., 2019), which complicates matters. It is obvious that inappropriate data aggregation creates temporal misalignment between different datasets. This in turn may artificially amplify forecast errors (see Fig. 1 in Yang, 2018). To select the correct data-aggregation scheme, it is necessary to understand how each dataset is produced. In other words, one must always read the data documentation. For instance, most radiometric networks use the ceiling convention, with the notable exception of the Baseline Surface Radiation Network (BSRN), which uses the floor convention. Satellite-derived irradiance and some NWP outputs have a "snapshot" nature, and the round-aggregation scheme is appropriate. In the case of reanalysis, the data often represent the condition over the past hour, and the ceiling-aggregation schemes is appropriate.

4.5. *Irradiance to PV power output conversion issues*

As discussed at the beginning, solar forecasting refers to both solar irradiance forecasting and solar power output forecasting. Given the fact that most modern solar forecasting methods leverage heavily on exogenous input, such as camera, satellite, or NWP data, time series methods are less frequently published (Yang, 2019a). In this regard, most researchers would take a two-step procedure to forecast solar power output: (1) forecast solar irradiance, and (2) convert the irradiance forecasts to power forecasts. For a fair comparison of forecasts, specific plant attributes need to be considered.

In the case of PV power plants, the panels are usually installed on a surface tilted at an angle close to the site's latitude, or with single- or double-axis trackers, to maximize the annual electricity production. Consequently, the variability of solar irradiance on the plane of array (POA) is often higher than that on a horizontal surface, due to the larger swing of irradiance. On the other hand, the physical size of the PV power plant should also be considered, because a PV plant behaves as a low-pass filter whose cut-off frequency depends on its size. In addition to this, due to the well-known geographical smoothing effect (Lave et al., 2012), the total output of a utility-scale PV power plant or a cluster of PV plants has less variability. Generally, areal forecasts are more accurate than the forecasts made for a point location (Yang et al., 2017). The reader is referred to Lohmann (2018) for a short review on spatio-temporal variability of solar irradiance, and to Perpiñán et al. (2013), Marcos et al. (2012) for analyses on power fluctuations in a large PV plant and a cluster of PV plants, respectively.

To convert the irradiance into PV power, three classes of models are often used, namely, separation models (Gueymard and Ruiz-Arias, 2016), transposition models (Yang, 2016), and irradiance-to-power

[18] http://www.soda-pro.com/web-services/radiation/cams-mcclear.

[19] This two-day lag makes it unsuitable for operational forecast verification. Nonetheless, during forecast verification, one is only interested in analyzing the long-term behavior of different models, so that this lag may not be of concern.

[20] This resolution has changed over time. For example, the resolution of aerosol optical depth has changed from 1.125° in 2004 to 0.4° in 2016.

conversion models (see Fig. 1 in Li et al., 2017). Separation models seek to estimate the DNI and DHI components from GHI; they are only needed if both DNI and DHI are unavailable. In the recent work by Yang and Boland (2019), several state-of-the-art separation models were compared, and the Yang2 model was shown to be one of the most accurate separation models. This result was later confirmed by Blaga (2019). On the other hand, transposition models convert horizontal irradiance components to the POA irradiances, and the Perez1990 model (Perez et al., 1990) is commonly regarded as one of the best-performing transposition models (Yang, 2016). Lastly, based on the POA irradiance and other parameters such as ambient or module temperature, the power output from a PV plant with known attributes can be predicted fairly accurately.

Although these modeling steps usually have a much higher accuracy than forecasting, their uncertainties have to be considered during forecast verification. Generally, due to cancellation of errors, the overall error of PV power forecasting is almost always smaller than the algebraic sum of errors in each step (forecasting, separation, transposition, and irradiance-to-power conversion). It is therefore of interest to study such error propagation, see Urraca et al. (2018), Almeida et al. (2017) for some preliminary findings. With the verification framework discussed earlier, the Murphy–Winkler factorizations could be applied to each of the steps, and thus provide new insights to solar engineers.

Lastly, it is worth noting that the baseline method for irradiance-to-power conversion is widely available in software packages, such as the System Advisor Model, PVSyst, or pvlib. These software packages take irradiance time series (e.g., typical meteorological year data) as input, and output PV power time series using the system specifications provided by users. For forecast verification purposes, one can feed the forecast irradiance time series, and obtain the corresponding forecast power time series. That said, emerging issues, such as lack of spectral and angular details in irradiance input (Lindsay et al., 2020) or the effect of cloud-enhancement phenomenon on PV power plants (Järvelä et al., 2020), are not yet being considered in these software packages, leaving room for improvements.

5. Conclusion

The increasing amount of solar forecasting research calls for harmonization of forecast verification measures and methods among researchers. This paper has discussed a wide spectrum of issues relevant to verification of deterministic solar forecasts. The final recommendations are listed as follows.

- The distribution-oriented approach to forecast verification can be used for forecast analysis. Since the joint distribution contains all time-independent information relevant to verification, it is more general than the traditional measure-oriented approach. It is recommended to use the distribution-oriented approach to visualize and quantify forecast quality.
- Bias–variance factorization and Murphy–Winkler factorizations link various qualitative aspects of the skillfulness of the forecasts (such as uncertainty, reliability, resolution, association, or discrimination) to quantitative measures. These decomposed measures provide forecasters with realistic and insightful assessments of the forecast verification problem, which essentially has a multifaceted nature.
- Small MBE is a prerequisite of all solar forecasts and therefore not a critical metric to judge forecast quality.
- When the normalized errors are reported, it is necessary to also tabulate the normalization values, such as the mean irradiance values.
- Generally, forecasters are encouraged to use any meaningful measure to gauge their forecasts. However, if a chosen measure is inconsistent with the given forecast directive, it might be inappropriate in theory.
- The RMSE skill score based on the optimal convex combination of single-valued internal climatology and h-step-ahead clear-sky persistence is strongly recommended in all solar forecasting studies. The skill score denotes the relative improvement of a model of interest from the reference method, and it can be used to compare forecasts produced in different works. Nonetheless, the skill score only provides an overall idea, so that the distribution-oriented approach is still required.
- Implementation issues are important for the final interpretation of forecast accuracy. Nighttime data should be excluded from forecast verification. Special care is needed during data trimming and aggregation, as a result of the necessary quality control and data gap-filling processes.
- As forecasting workflows are getting more and more complex, it is advised to perform sanity checks throughout the course of producing forecasts. To ensure the worldwide uptake of any proposed forecasting model, source code and data should be made available whenever possible. Without reproducibility, it would be cumbersome—if not impossible—to verify the reported forecast performance.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The lead author would like to thank Cong Feng (University of Texas at Dallas), Bri-Mathias Hodge (National Renewable Energy Laboratory), and Pierre Pinson (Technical University of Denmark) for their discussions during manuscript preparation. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Appendix A. On seasonal adjustment of reference forecasts

In the simplest case, persistence, by definition, uses the most recent observation available as forecasts for all horizons. Such forecast is also sometimes referred to as “random walk” (i.e., $r_t = r_{t-1} + e_t$), or “no change” forecast. In the case of solar irradiance forecasting, raw persistence forecasts should be the most recent observed irradiance. Notwithstanding, given the bell-shaped diurnal transient of irradiance caused by the Sun’s apparent movement, it is important to take such seasonality into consideration. Makridakis et al. (2008) noted that seasonally adjusted persistence can frequently do much better than the raw persistence. So the question is “how is it best to adjust for seasonality?”.

Besides using the clear-sky irradiance, one can use the extraterrestrial irradiance (the irradiance at the top of the atmosphere) for adjustment. The ratio between surface and extraterrestrial GHI is known as clearness index, k. In other words, the persistence is performed on clearness index, namely, $r_{t+h} = x_t \cdot k_{t+h}/k_t$. Both clear-sky persistence and clearness persistence adopt a multiplicative seasonality modeling approach. A particular problem with multiplicative seasonality is that during sunrise and sunset (small solar elevation angle), both the clear-sky index and clearness index can become quite large, owing to the measurement uncertainty and the inaccuracy in the clear-sky models, and thus the forecast errors can be large at those times. To exclude those undesirable forecasts that may severely distort the error metrics, solar forecasters usually apply a zenith angle filter, e.g., zenith angle <85°, before error computation. Alternatively, one can opt for an additive seasonality modeling, e.g., $r_{t+h} = x_t - c_t + c_{t+h}$. However, the remainder series (i.e., $x_t - c_t$) in this case is still heteroscedastic.

Alternatively, one can also use a "cloudiness index" where the reference forecast is referred to as "smart persistence" by Inman et al. (2015). This reference model includes the effects of air mass, aerosols, turbidity, i.e., every major atmospheric effect except that of clouds. Because the timescale for turbidity variations is much larger than the timescale for cloud-cover variations, the cloudiness index provides an excellent reference for short-time forecasts. Any skill calculated over the cloudiness-index persistence measures the ability to capture cloud cover changes over short periods of time, and thus the qualifier "smart". Note that skills reported over smart persistence are necessarily lower than skills reported over clear-sky persistence. This is because it is virtually impossible for a forecast to improve over smart persistence under cloudless skies, since the smart persistence reference model includes all effects of diurnal variability (solar zenith angle) and columnar optical depths of water vapor and aerosols. These types of smart persistence reference forecasts are typically updated sub-hourly (Inman et al., 2015; Reno and Hansen, 2016).

The literature on the treatment of seasonal and multi-seasonal (e.g., diurnal and yearly cycles in solar irradiance) time series is rich. Chapter 3 of Makridakis et al. (2008) provides details on various statistical techniques for time series decomposition. In the solar forecasting literature, various techniques, such as Fourier series, exponential smoothing, STL decomposition, additive clear-sky decomposition, or smoothed clear-sky decomposition, have also been extensively explored (e.g., Dong et al., 2013; Yang et al., 2015; Voyant and Notton, 2018). However, as compared to the clear-sky persistence, these methods usually require more steps, which may be a reason for their limited uptake. Since the main goal of seasonally-adjusted persistence is to construct a better reference model than raw persistence, clear-sky persistence offers a good trade-off between implementation difficulty and baseline accuracy.

Appendix B. On equivalence of s^* and s

Denoting the clear-sky index forecasts of interest, reference forecasts, and observations using ϕ, ρ, and $\varkappa$, respectively, one can write: $\phi_t \equiv f_t/c_t$, $\rho_t \equiv r_t/c_t$, and $\varkappa_t \equiv x_t/c_t$. For 1-step-ahead clear-sky persistence, $r_t = x_{t-1}\cdot c_t/c_{t-1}$, hence, $\rho_t = x_{t-1}/c_{t-1}$. Eqs. (8) and (9) can then be written as:

$$U = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\phi_t - \varkappa_t)^2} = \mathrm{RMSE}(\phi, \varkappa), \tag{B.1}$$

$$V = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(\rho_t - \varkappa_t)^2} = \mathrm{RMSE}(\rho, \varkappa). \tag{B.2}$$

Therefore, $s^* = 1 - \mathrm{RMSE}(\phi, \varkappa)/\mathrm{RMSE}(\rho, \varkappa)$ is the RMSE skill score of clear-sky index forecasts, with 1-step-ahead persistence reference.

Suppose a perfect clear-sky model exists, which explains all influences of atmospheric constituents but those from clouds, then one can say that the clear-sky index is independent of the clear-sky expectation, i.e., $\varkappa \perp\!\!\!\perp c$. Hence, any forecast error, $e = \phi - \varkappa = g(\varkappa) - \varkappa$, where g is a function to produce $\varkappa$ forecast, is also independent of c, i.e,. $e \perp\!\!\!\perp c$. From the properties of independence, it immediately follows that:

$$\mathbb{E}(e^2c^2) = \mathbb{E}(c^2)\mathbb{E}(e^2). \tag{B.3}$$

That is:

$$\mathrm{MSE}(f, x) = \mathbb{E}(c^2)\mathrm{MSE}(\phi, \varkappa). \tag{B.4}$$

Stated differently, if $\varkappa \perp\!\!\!\perp c$, the MSE of clear-sky index forecasts and that of irradiance forecasts are scaled by a factor of $\mathbb{E}(c^2)$. Equivalently,

$$\mathrm{RMSE}(f, x) = \sqrt{\mathbb{E}(c^2)}\cdot\mathrm{RMSE}(\phi, \varkappa). \tag{B.5}$$

Since g is any function, the result also applies to the reference $\varkappa$ forecast. Then,

$$s^* = 1 - \frac{\mathrm{RMSE}(\phi, \varkappa)}{\mathrm{RMSE}(\rho, \varkappa)} = 1 - \frac{\sqrt{\mathbb{E}(c^2)}\cdot\mathrm{RMSE}(\phi, \varkappa)}{\sqrt{\mathbb{E}(c^2)}\cdot\mathrm{RMSE}(\rho, \varkappa)} = 1 - \frac{\mathrm{RMSE}(f, x)}{\mathrm{RMSE}(r, x)} = s. \tag{B.6}$$

That said, it should be clearly noted that the assumption $\varkappa \perp\!\!\!\perp c$ is *not* valid in all cases (Yang, 2020). This is because even the best clear-sky models could not completely explain the atmospheric effects on radiation. Hence, when there is some degree of dependence between $\varkappa$ and c, one can only observe $s^* \approx s$, as empirically shown by Marquez and Coimbra (2011, 2013), Coimbra et al. (2013). This issue deserves future attention; the reader is referred to Yang (2020) for a discussion on stationarity of $\varkappa$.

Appendix C. Links between measure-oriented and distribution-oriented forecast verification approaches

The *rule of the lazy statistician* states (Wasserman, 2013): Let $y = g(x)$, then

$$\mathbb{E}(y) = \mathbb{E}[g(x)\bigg] = \int g(x)dF(x) = \int g(x)p(x)dx. \tag{C.1}$$

The two-variable case is handled in a similar way: Let $z = g(x, y)$, then

$$\mathbb{E}(z) = \mathbb{E}[g(x, y)] = \int g(x, y)dF(x, y) = \iint g(x, y)p(x, y)dxdy. \tag{C.2}$$

This rule links the joint distribution to a large collection of error metrics. For example, MBE, MAE, and RMSE can be written as:

$$\mathrm{MBE} = \mathbb{E}[(f - x)] = \iint (f - x)p(f, x)dfdx, \tag{C.3}$$

$$\mathrm{MAE} = \mathbb{E}(|f - x|) = \iint |f - x|p(f, x)dfdx, \tag{C.4}$$

$$\text{RMSE} = \sqrt{\mathbb{E}\left[(f-x)^2\right]} = \left[\iint (f-x)^2 p(f,x)\,df\,dx\right]^{\frac{1}{2}}. \tag{C.5}$$

Similarly, it is possible to express nMBE, nMAE, nRMSE, maximum absolute error, mean average percentage error, etc., in this form. Hence, it is clear that all these metrics are just different ways of summarizing the joint distribution.

In the report by Beyer et al. (2009), four metrics based on the Kolmogrov–Smirnov test were proposed, namely, the Kolmogrov–Smirnov test integral (KSI), the OVER index, KSE (linear combination of KSI and OVER), and RIO (sum of KSD and RMSE, divided by 2). For instance, KSI calculates approximately the area between the ECDFs of f and x, whereas OVER calculates the area of those instances between the two ECDFs that exceed the critical value at 99% level of confidence. While the definitions of these metrics might deviate from the usual statistical measures (such as the Wasserstein distance), they actually constitute an early attempt by solar engineers to summarize the differences between marginal distributions of f and x.

Lastly, Section 3 demonstrates several possible ways of summarizing the conditional distributions. In these cases, the summaries would also require marginal distributions. More specifically, the type-1 conditional bias and resolution are summaries of $p(x|f)$ and $p(f)$, whereas the type-2 conditional bias and discrimination are summaries of $p(f|x)$ and $p(x)$ (Murphy, 1997).

The remaining open questions, such as “are there better ways of summarizing these distributions,” “which summaries allow cross-work forecast comparison,” “how to analyze the summaries graphically,” etc., jointly motivate future research on verification of deterministic solar forecasts.

Appendix D. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.solener.2020.04.019.

References

Almeida, M.P., Muñoz, M., de la Parra, I., Perpiñán, O., 2017. Comparative study of PV power forecast using parametric and nonparametric PV models. Sol. Energy 155, 854–866. https://doi.org/10.1016/j.solener.2017.07.032.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X17306175.

Antonanzas, J., Osorio, N., Escobar, R., Urraca, R., de Pison, F.M., Antonanzas-Torres, F., 2016. Review of photovoltaic power forecasting. Sol. Energy 136, 78–111. https://doi.org/10.1016/j.solener.2016.06.069.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X1630250X.

Antonanzas, J., Pozo-Vázquez, D., Fernandez-Jimenez, L., de Pison, F.M., 2017. The value of day-ahead forecasting for photovoltaics in the Spanish electricity market. Sol. Energy 158, 140–146. https://doi.org/10.1016/j.solener.2017.09.043.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X17308307.

Antonanzas-Torres, F., Urraca, R., Polo, J., Perpiñán-Lamigueiro, O., Escobar, R., 2019. Clear sky solar irradiance models: A review of seventy models. Renew. Sustain. Energy Rev. 107, 374–387. https://doi.org/10.1016/j.rser.2019.02.032.. URL: http://www.sciencedirect.com/science/article/pii/S1364032119301261.

Armstrong, J.S., 2001. Evaluating forecasting methods. In: Principles of Forecasting. Springer, pp. 443–472.

Beyer, H.G., Polo Martinez, J., Suri, M., Torres, J.L., Lorenz, E., Müller, S.C., Hoyer-Klick, C., Ineichen, P., 2009. Benchmarking of Radiation Products. Technical Report 038665. Mesor Report D.1.1.3.

Blaga, R., 2019. The impact of temporal smoothing on the accuracy of separation models. Sol. Energy 191, 371–381. https://doi.org/10.1016/j.solener.2019.08.078.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X19308709.

Blaga, R., Sabadus, A., Stefu, N., Dughir, C., Paulescu, M., Badescu, V., 2019. A current perspective on the accuracy of incoming solar energy forecasting. Prog. Energy Combust. Sci. 70, 119–144. https://doi.org/10.1016/j.pecs.2018.10.003.. URL: http://www.sciencedirect.com/science/article/pii/S0360128518300303.

Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geoscientific Model Devel. 7, 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014.. URL: https://www.geosci-model-dev.net/7/1247/2014/.

Coimbra, C.F.M., Kleissl, J., Marquez, R., 2013. Chapter 8 - Overview of solar-forecasting methods and a metric for accuracy evaluation. In: Kleissl, J. (Ed.), Solar Energy Forecasting and Resource Assessment. Academic Press, Boston, pp. 171–194. https://doi.org/10.1016/B978-0-12-397177-7.00008-5.. URL: http://www.sciencedirect.com/science/article/pii/B9780123971777000085.

Dong, Z., Yang, D., Reindl, T., Walsh, W.M., 2013. Short-term solar irradiance forecasting using exponential smoothing state space model. Energy 55, 1104–1113. https://doi.org/10.1016/j.energy.2013.04.027.. URL: http://www.sciencedirect.com/science/article/pii/S0360544213003381.

Fildes, R., Nikolopoulos, K., Crone, S.F., Syntetos, A.A., 2008. Forecasting and operational research: a review. J. Oper. Res. Soc. 59, 1150–1172. https://doi.org/10.1057/palgrave.jors.2602597.

García, R.D., García, O.E., Cuevas, E., Cachorro, V.E., Romero-Campos, P.M., Ramos, R., de Frutos, A.M., 2014. Solar radiation measurements compared to simulations at the BSRN Izaña station. mineral dust radiative forcing and efficiency study. J. Geophys. Res.: Atmosph. 119, 179–194. https://doi.org/10.1002/2013JD020301.. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2013JD020301.

Gilleland, E., Ahijevych, D.A., Brown, B.G., Ebert, E.E., 2010. Verifying forecasts spatially. Bull. Am. Meteorol. Soc. 91, 1365–1376. https://doi.org/10.1175/2010BAMS2819.1.

Gneiting, T., 2011. Making and evaluating point forecasts. J. Am. Stat. Assoc. 106, 746–762. https://doi.org/10.2307/41416407.. URL: http://www.jstor.org/stable/41416407.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Am. Statist. Assoc. 102, 359–378. https://doi.org/10.1198/016214506000001437.

Gschwind, B., Wald, L., Blanc, P., Lefèvre, M., Schroedter-Homscheidt, M., Arola, A., 2019. Improving the McClear model estimating the downwelling solar radiation at ground level in cloud-free conditions – McClear-v3. Meteorol. Z. 28, 147–163. https://doi.org/10.1127/metz/2019/0946.

Gueymard, C.A., 2012. Clear-sky irradiance predictions for solar resource mapping and large-scale applications: Improved validation methodology and detailed performance analysis of 18 broadband radiative models. Sol. Energy 86, 2145–2169. URL: http://www.sciencedirect.com/science/article/pii/S0038092X11004221. https://doi.org/10.1016/j.solener.2011.11.011. Progress in Solar Energy 3.

Gueymard, C.A., Ruiz-Arias, J.A., 2016. Extensive worldwide validation and climate sensitivity analysis of direct irradiance predictions from 1-min global irradiance. Sol. Energy 128, 1–30. URL: http://www.sciencedirect.com/science/article/pii/S0038092X15005435. https://doi.org/10.1016/j.solener.2015.10.010. Special issue: Progress in Solar Energy.

Hoff, T.E., Perez, R., Kleissl, J., Renne, D., Stein, J., 2013. Reporting of irradiance modeling relative prediction errors. Prog. Photovoltaics Res. Appl. 21, 1514–1519. https://doi.org/10.1002/pip.2225.. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pip.2225.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., Hyndman, R.J., 2016. Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond. Int. J. Forecast. 32, 896–913. https://doi.org/10.1016/j.ijforecast.2016.02.001.. URL: http://www.sciencedirect.com/science/article/pii/S0169207016000133.

Huang, J., Thatcher, M., 2017. Assessing the value of simulated regional weather variability in solar forecasting using numerical weather prediction. Sol. Energy 144, 529–539. https://doi.org/10.1016/j.solener.2017.01.058.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X17300774.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. Int. J. Forecast. 22, 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001.. URL: http://www.sciencedirect.com/science/article/pii/S0169207006000239.

Inman, R.H., Edson, J.G., Coimbra, C.F.M., 2015. Impact of local broadband turbidity estimation on forecasting of clear sky direct normal irradiance. Sol. Energy 117, 125–138. https://doi.org/10.1016/j.solener.2015.04.032.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X15002200.

Inman, R.H., Pedro, H.T.C., Coimbra, C.F.M., 2013. Solar forecasting methods for renewable energy integration. Prog. Energy Combust. Sci. 39, 535–576. https://doi.org/10.1016/j.pecs.2013.06.002.. URL: http://www.sciencedirect.com/science/article/pii/S0360128513000294.

Järvelä, M., Lappalainen, K., Valkealahti, S., 2020. Characteristics of the cloud enhancement phenomenon and PV power plants. Sol. Energy 196, 137–145. https://doi.org/10.1016/j.solener.2019.11.090.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X19311909.

Jolliffe, I.T., 2008. The impenetrable hedge: a note on propriety, equitability and consistency. Meteorolog. Appl. 15, 25–29. https://doi.org/10.1002/met.60.. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.60.

Jolliffe, I.T., Stephenson, D.B., 2012. Forecast Verification: A Practitioner’s Guide in Atmospheric Science. John Wiley & Sons.

Killinger, S., Engerer, N., Müller, B., 2017. QCPV: A quality control algorithm for distributed photovoltaic array power output. Sol. Energy 143, 120–131.

Klingler, A.L., Teichtmann, L., 2017. Impacts of a forecast-based operation strategy for grid-connected PV storage systems on profitability and the energy system. Sol. Energy 158, 861–868. https://doi.org/10.1016/j.solener.2017.10.052.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X17309325.

Lave, M., Kleissl, J., Arias-Castro, E., 2012. High-frequency irradiance fluctuations and geographic smoothing. Sol. Energy 86, 2190–2199. URL: http://www.sciencedirect.com/science/article/pii/S0038092X11002611. https://doi.org/10.1016/j.solener.

2011.06.031. Progress in Solar Energy 3.

Law, E.W., Kay, M., Taylor, R.A., 2016. Calculating the financial value of a concentrated solar thermal plant operated using direct normal irradiance forecasts. Sol. Energy 125, 267–281. https://doi.org/10.1016/j.solener.2015.12.031.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X15007045.

Lefèvre, M., Oumbe, A., Blanc, P., Espinar, B., Gschwind, B., Qu, Z., Wald, L., Schroedter-Homscheidt, M., Hoyer-Klick, C., Arola, A., Benedetti, A., Kaiser, J.W., Morcrette, J.J., 2013. McClear: a new model estimating downwelling solar radiation at ground level in clear-sky conditions. Atmospheric Measur. Tech. 6, 2403–2418. https://doi.org/10.5194/amt-6-2403-2013.. URL: https://www.atmos-meas-tech.net/6/2403/2013/.

Li, Y., Chen, X., Zhao, B., Zhao, Z., Wang, R., 2017. Development of a PV performance model for power output simulation at minutely resolution. Renewable Energy 111, 732–739. https://doi.org/10.1016/j.renene.2017.04.049.. URL: http://www.sciencedirect.com/science/article/pii/S0960148117303567.

Lindsay, N., Libois, Q., Badosa, J., Migan-Dubois, A., Bourdin, V., 2020. Errors in PV power modelling due to the lack of spectral and angular details of solar irradiance inputs. Sol. Energy 197, 266–278. https://doi.org/10.1016/j.solener.2019.12.042.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X19312563.

Lohmann, G.M., 2018. Irradiance variability quantification and small-scale averaging in space and time: A short review. Atmosphere 9https://doi.org/10.3390/atmos9070264.. URL: https://www.mdpi.com/2073-4433/9/7/264.

Long, C.N., Shi, Y., 2008. An automated quality assessment and control algorithm for surface radiation measurements. Open Atmospheric Sci. J. 2, 23–37.

Lorenz, E., Kühnert, J., Heinemann, D., Nielsen, K.P., Remund, J., Müller, S.C., 2016. Comparison of global horizontal irradiance forecasts based on numerical weather prediction models with different spatio-temporal resolutions. Prog. Photovoltaics Res. Appl. 24, 1626–1640. https://doi.org/10.1002/pip.2799.. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pip.2799.

Madsen, H., Pinson, P., Kariniotakis, G., Nielsen, H.A., Nielsen, T.S., 2005. Standardizing the performance evaluation of short-term wind power prediction models. Wind Eng. 29, 475–489. https://doi.org/10.1260/030952405776234599.

Makridakis, S., Wheelwright, S.C., Hyndman, R.J., 2008. Forecasting Methods and Applications. John Wiley & Sons.

Marcos, J., Marroyo, L., Lorenzo, E., Garcáa, M., 2012. Smoothing of PV power fluctuations by geographical dispersion. Prog. Photovoltaics Res. Appl. 20, 226–237. https://doi.org/10.1002/pip.1127.. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pip.1127.

Marquez, R., Coimbra, C.F.M., 2011. A novel metric for evaluation of solar forecasting models. In: ASME 2011 5th International Conference on Energy Sustainability. ASME, pp. 1459–1467. https://doi.org/10.1115/ES2011-54519.

Marquez, R., Coimbra, C.F.M., 2013. Proposed metric for evaluation of solar forecasting models. J. Solar Energy Eng. 135, 011016. https://doi.org/10.1115/1.4007496.

Martinez-Anido, C.B., Botor, B., Florita, A.R., Draxl, C., Lu, S., Hamann, H.F., Hodge, B.M., 2016. The value of day-ahead solar power forecasting improvement. Sol. Energy 129, 192–203. https://doi.org/10.1016/j.solener.2016.01.049.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X16000736.

van der Meer, D., Widén, J., Munkhammar, J., 2018. Review on probabilistic forecasting of photovoltaic power production and electricity consumption. Renew. Sustain. Energy Rev. 81, 1484–1512. https://doi.org/10.1016/j.rser.2017.05.212.. URL: http://www.sciencedirect.com/science/article/pii/S1364032117308523.

Moskaitis, J.R., 2008. A case study of deterministic forecast verification: Tropical cyclone intensity. Weather Forecast. 23, 1195–1220. https://doi.org/10.1175/2008WAF2222133.1.

Murphy, A.H., 1988. Skill scores based on the mean square error and their relationships to the correlation coefficient. Mon. Weather Rev. 116, 2417–2424. https://doi.org/10.1175/1520-0493(1988)116<2417:SSBOTM>2.0.CO;2.

Murphy, A.H., 1992. Climatology, persistence, and their linear combination as standards of reference in skill scores. Weather Forecast. 7, 692–698. https://doi.org/10.1175/1520-0434(1992)007<0692:CPATLC>2.0.CO;2.

Murphy, A.H., 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. Weather Forecast. 8, 281–293. https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Murphy, A.H., 1997. Forecast verification. In: Katz, R.W., Murphy, A.H. (Eds.), Economic Value of Weather and Climate Forecasts. Cambridge University Press, pp. 19–74. https://doi.org/10.1017/CBO9780511608278.003.

Murphy, A.H., Brown, B.G., Chen, Y.S., 1989. Diagnostic verification of temperature forecasts. Weather Forecast. 4, 485–501. https://doi.org/10.1175/1520-0434(1989)004<0485:DVOTF>2.0.CO;2.

Murphy, A.H., Winkler, R.L., 1971. forecasters and probability forecasts: some current problems. Bull. Am. Meteorol. Soc. 52, 239–248. https://doi.org/10.1175/1520-0477(1971)052<0239:FAPFSC>2.0.CO;2.

Murphy, A.H., Winkler, R.L., 1987. A general framework for forecast verification. Mon. Weather Rev. 115, 1330–1338. https://doi.org/10.1175/1520-0493(1987)115<1330:AGFFFV>2.0.CO;2.

Pedro, H.T.C., Coimbra, C.F.M., 2015. Short-term irradiance forecastability for various solar micro-climates. Sol. Energy 122, 587–602. https://doi.org/10.1016/j.solener.2015.09.031.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X15005162.

Perez, R., Ineichen, P., Seals, R., Michalsky, J., Stewart, R., 1990. Modeling daylight availability and irradiance components from direct and global irradiance. Sol. Energy 44, 271–289. https://doi.org/10.1016/0038-092X(90)90055-H.. URL: http://www.sciencedirect.com/science/article/pii/0038092X9090055H.

Perez, R., Lorenz, E., Pelland, S., Beauharnois, M., Knowe, G.V., Hemker, K., Heinemann, D., Remund, J., Müller, S.C., Traunmüller, W., Steinmauer, G., Pozo, D., Ruiz-Arias, J.A., Lara-Fanego, V., Ramirez-Santigosa, L., Gaston-Romero, M., Pomares, L.M., 2013. Comparison of numerical weather prediction solar irradiance forecasts in the US, Canada and Europe. Sol. Energy 94, 305–326. https://doi.org/10.1016/j.solener.2013.05.005.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X13001886.

Perez, R., Schlemmer, J., Kankiewicz, A., Dise, J., Tadese, A., Hoff, T., 2017. Detecting calibration drift at ground truth stations a demonstration of satellite irradiance models' accuracy. In: 2017 IEEE 44th Photovoltaic Specialist Conference (PVSC), pp. 1104–1109. https://doi.org/10.1109/PVSC.2017.8366469.

Perpiñán, O., Marcos, J., Lorenzo, E., 2013. Electrical power fluctuations in a network of DC/AC inverters in a large PV plant: Relationship between correlation, distance and time scale. Sol. Energy 88, 227–241. https://doi.org/10.1016/j.solener.2012.12.004.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X12004197.

Polo, J., Martín-Pomares, L., Gueymard, C.A., Balenzategui, J.L., Fabero, F., Silva, J.P., 2019. Fundamentals: Quantities, definitions, and units. In: Polo, J., Martín-Pomares, L., Sanfilippo, A. (Eds.), Solar Resources Mapping: Fundamentals and Applications. Springer International Publishing, Cham, pp. 1–14. https://doi.org/10.1007/978-3-319-97484-2_1.

Ren, Y., Suganthan, P., Srikanth, N., 2015. Ensemble methods for wind and solar power forecasting—A state-of-the-art review. Renew. Sustain. Energy Rev. 50, 82–91. https://doi.org/10.1016/j.rser.2015.04.081.. URL: http://www.sciencedirect.com/science/article/pii/S1364032115003512.

Reno, M.J., Hansen, C.W., 2016. Global horizontal irradiance clear sky models: Implementation and analysis. Renewable Energy 90, 520–531. https://doi.org/10.1016/j.renene. 938 2015.12.031.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X15002200.

Ruiz-Arias, J.A., Gueymard, C.A., 2018b. Worldwide inter-comparison of clear-sky solar radiation models: Consensus-based review of direct and global irradiance components simulated at the earth surface. Sol. Energy 168, 10–29. Advances in Solar Resource Assessment and Forecasting. URL: http://www.sciencedirect.com/science/article/pii/S0038092X18301257. https://doi.org/10.1016/j.solener.2018.02.008.

Ruiz-Arias, J.A., Gueymard, C.A., 2018a. A multi-model benchmarking of direct and global clear-sky solar irradiance predictions at arid sites using a reference physical radiative transfer model. Sol. Energy 171, 447–465. https://doi.org/10.1016/j.solener.2018.06.048.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X18305991.

Schilling, R.L., 2017. Measures, Integrals and Martingales. Cambridge University Press.

Sengupta, M., Habte, A., Kurtz, S., Dobos, A., Wilbert, S., Lorenz, E., Stoffel, T., Renné, D., Gueymard, C.A., Myers, D., et al., 2015. Best practices handbook for the collection and use of solar resource data for solar energy applications. Technical Report NREL/TP-5D00-63112. National Renewable Energy Laboratory.

Sun, X., Bright, J.M., Gueymard, C.A., Acord, B., Wang, P., Engerer, N.A., 2019. Worldwide performance assessment of 75 global clear-sky irradiance models using principal component analysis. Renew. Sustain. Energy Rev. 111, 550–570. https://doi.org/10.1016/j.rser.2019.04.006.. URL: http://www.sciencedirect.com/science/article/pii/S1364032119302187.

Tashman, L.J., 2000. Out-of-sample tests of forecasting accuracy: an analysis and review. Int. J. Forecast. 16, 437–450. URL: http://www.sciencedirect.com/science/article/pii/S0169207000000650. https://doi.org/10.1016/S0169-2070(00)00065-0. The M3- Competition.

Urraca, R., Gracia-Amillo, A.M., Huld, T., de Pison, F.J.M., Trentmann, J., Lindfors, A.V., Riihelä, A., Sanz-Garcia, A., 2017. Quality control of global solar radiation data with satellite-based products. Sol. Energy 158, 49–62.

Urraca, R., Huld, T., Lindfors, A.V., Riihelä, A., de Pison, F.J.M., Sanz-Garcia, A., 2018. Quantifying the amplified bias of PV system simulations due to uncertainties in solar radiation estimates. Sol. Energy 176, 663–677. https://doi.org/10.1016/j.solener.2018.10.065.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X18310442.

Vallance, L., Charbonnier, B., Paul, N., Dubost, S., Blanc, P., 2017. Towards a standardized procedure to assess solar forecast accuracy: A new ramp and time alignment metric. Sol. Energy 150, 408–422. https://doi.org/10.1016/j.solener.2017.04.064.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X17303687.

Voyant, C., Notton, G., 2018. Solar irradiation nowcasting by stochastic persistence: A new parsimonious, simple and efficient forecasting tool. Renew. Sustain. Energy Rev. 92, 343–352. https://doi.org/10.1016/j.rser.2018.04.116.. URL: http://www.sciencedirect.com/science/article/pii/S1364032118303344.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M.L., Paoli, C., Motte, F., Fouilloy, A., 2017. Machine learning methods for solar radiation forecasting: A review. Renewable Energy 105, 569–582. https://doi.org/10.1016/j.renene.2016.12.095.. URL: http://www.sciencedirect.com/science/article/pii/S0960148116311648.

Wasserman, L., 2013. All of Statistics: A Concise Course in Statistical Inference. Springer Science & Business Media.

Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Res. 30, 79–82. https://doi.org/10.3354/cr030079.. URL: https://www.int-res.com/abstracts/cr/v30/n1/p79-82/.

Yang, D., 2016. Solar radiation on inclined surfaces: Corrections and benchmarks. Sol. Energy 136, 288–302. https://doi.org/10.1016/j.solener.2016.06.062.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X16302432.

Yang, D., 2018. A correct validation of the National Solar Radiation Data Base (NSRDB). Renew. Sustain. Energy Rev. 97, 152–155. https://doi.org/10.1016/j.rser.2018.08.023.. URL: http://www.sciencedirect.com/science/article/pii/S1364032118306087.

Yang, D., 2019a. A guideline to solar forecasting research practice: Reproducible, operational, probabilistic or physically-based, ensemble, and skill (ROPES). J. Renewable Sustainable Energy 11, 22701. https://doi.org/10.1063/1.5087462.

Yang, D., 2019b. Making reference solar forecasts with climatology, persistence, and their optimal convex combination. Sol. Energy 193, 981–985. https://doi.org/10.1016/j.

solener.2019.10.006.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X19309880.

Yang, D., 2019c. Standard of reference in operational day-ahead deterministic solar forecasting. J. Renewable Sustainable Energy 11, 53702. https://doi.org/10.1063/1.5114985.

Yang, D., 2020. Choice of clear-sky model in solar forecasting. J. Renewable Sustainable Energy 12 (2), 26101. https://doi.org/10.1063/5.0003495.

Yang, D., Boland, J., 2019. Satellite-augmented diffuse solar radiation separation models. J. Renewable Sustainable Energy 11, 023705. https://doi.org/10.1063/1.5087463.

Yang, D., Kleissl, J., Gueymard, C.A., Pedro, H.T.C., Coimbra, C.F.M., 2018. History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining. Sol. Energy 168, 60–101. Advances in Solar Resource Assessment and Forecasting. URL: http://www.sciencedirect.com/science/article/pii/S0038092X17310022. https://doi.org/10.1016/j.solener.2017.11.023.

Yang, D., Perez, R., 2019. Can we gauge forecasts using satellite-derived solar irradiance? J. Renewable Sustainable Energy 11, 023704. https://doi.org/10.1063/1.5087588.

Yang, D., Quan, H., Disfani, V.R., Liu, L., 2017. Reconciling solar forecasts: Geographical hierarchy. Sol. Energy 146, 276–286. https://doi.org/10.1016/j.solener.2017.02.010.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X17301020.

Yang, D., Sharma, V., Ye, Z., Lim, L.I., Zhao, L., Aryaputera, A.W., 2015. Forecasting of global horizontal irradiance by exponential smoothing, using decompositions. Energy 81, 111–119. https://doi.org/10.1016/j.energy.2014.11.082.. URL: http://www.sciencedirect.com/science/article/pii/S0360544214013528.

Zhang, J., Florita, A., Hodge, B.M., Lu, S., Hamann, H.F., Banunarayanan, V., Brockway, A.M., 2015. A suite of metrics for assessing the performance of solar power forecasting. Sol. Energy 111, 157–175. https://doi.org/10.1016/j.solener.2014.10.016.. URL: http://www.sciencedirect.com/science/article/pii/S0038092X14005027.